WPS Working Paper Series

Vol. I, No. 2, 2016

ENGLISH

Inter-American Conference on Social Security

WPS Working Paper Series

Risk-profiling of Potential Diabetics at IMSS A Logistic Regression Approach

Ari Bronsoler
Christian Norton
Óscar Sánchez
Carlos Tendilla*

Abstract

Modern public medicine is relying more and more on preventive rather than corrective action. This is happening because preventive care is proving to be not only cost-effective but also desirable, as it can reduce length of convalescence and treatment expenditures while allowing for better living conditions for patients and improving longevity. In this document we describe the methodological steps by which we are able to estimate the risk of being diagnosed with Type 2 Diabetes Mellitus on individuals that attended a medical clinic from Mexico's Institute for Social Security (IMSS) between 2012 and 2014. The results of this investigation lead to practical conclusions that show, for instance, that by applying our risk-profiling criteria for confirmatory laboratory test referral and without performing any additional medical tests, 50 thousand additional diabetes cases would have been detected, which means a 90% increase in diagnosis. Highlighting the public-policy relevance of these conclusions, and leveraging the structure of IMSS databases, we introduce a simple questionnaire that would allow risk-profiling to be applied to the population at large.

* The present document is the result of research undertaken at Mexico's Social Security Institute and would not have been possible without the support of the Institute's Director General José Antonio González. Dr. González not only believed in the project from the outset but offered his attention and feedback in the midst of having to attend to the day-to-day operation of one of Mexico's largest public institutions. The conclusions derived herein were also attainable thanks to having had access to vast databases at IMSS. Dr. Judith Frías, Eduardo Alcaraz and Samuel Trujillo –from the Incorporation and Revenue Directorate– provided guidance and the data on socio-economic characteristics of IMSS beneficiaries. César Hernández and Ángel Campos –from the medical infrastructure planning coordination– contributed with the infrastructure and medical information, and walked us through the data filtering process. Dr. Víctor Hugo Borja and his team at the Public Health Unit, as well as Dr. Alma Páez from the primary care medical wing, offered unconditional advice throughout the medical profiling exercise. Finally, systems engineer Javier Rodríguez–from the Technological Development Directorate–was instrumental in guiding us through the learning process on the status quo of computational systems within the Institute. We remain deeply indebted to each and every one of these individuals for their selfless guidance. The views herein expressed are solely our own and do not represent those held at IMSS. We are responsible for each and every error remaining in the document.

Working Paper Series
2016, vol. I, no. 2, pp. 133-183

Introduction

Between 2000 and 2012 life-expectancy in Mexico increased by less than a year, while in OECD countries it rose by three years on average. This implies that Mexico not only retains the lowest longevity within OECD, but also that the gap in life-expectancy with these countries continues to widen steadily. This dynamic is explained in part by the fact that, due to the country's demographic and epidemiological transition, Mexico's burden of chronic non-communicable diseases has dramatically increased in past decades. Consequently, the prevalence of these diseases is causing a significant fall in disability-adjusted life years (DALYs), with their incidence impacting the financial viability of the country's public health institutions. In order of importance the main health sufferings of Mexico's population today are: diabetes mellitus; hypertension; chronic kidney failure; cervical cancer; breast cancer and HIV.[1]

According to the International Diabetes Foundation's Atlas for 2014, Mexico ranked first amongst OECD countries in the number of diabetes cases per capita, as is displayed in figure 1.[2] In terms of mortality, deaths caused by the disease increased from 21.8 per 100,000 people in 1980, to 62 in 2011.[3] This is explained in part by the fact that the country ranks 2nd in the number of hospital admissions related with diabetes for each 100,000 people, as shown in figure 2. Incidence continues to creep up; with Mexico's National Health and Nutrition Survey (Ensanut) for 2012 identifying 6.4 million adults diagnosed with the disease and estimating double the prevalence if accounting for the non-diagnosed population. The same report calculates that 2.7 million IMSS beneficiaries lived with the pathology in 2012.[4]

Alongside the rest of Mexico's public health institutions, IMSS is tackling the situation head on as the Institute's bottom line is affected in a significant way by incidence from medical complications linked to diabetes. The study hereby introduced is part of an investigation undertaken by advisors to the Institute's Director General, and is based on an unprecedented effort to amalgamate socio-economic, medical and infrastructure data at the beneficiary level. The results are obtained by applying an in-house developed econometric algorithm, from which the most relevant result is a risk-screening for DM-2 of all IMSS beneficiaries.[5]

[1] Mexican Institute of Social Security's report to the executive branch and to Congress on its financial situation and its risks, 2013 – 2014. http://www.imss.gob.mx/sites/all/statics/pdf/informes/20132014/21_InformeCompleto.pdf.

[2] International Diabetes Foundation Atlas 2014. http://www.idf.org/diabetesatlas.

[3] Prevention of Overweight and Obesity Emerging Program, Mexico's Ministry of Health, 2011. http://www.salud.gob.mx/unidades/cdi/pot/fxi/CENAPRECE/PROG2011_2012.pdf.

[4] Encuesta Nacional de Salud y Nutrición 2012, Evidencia para política pública en salud, Diabetes Mellitus: la urgencia de reforzar la respuesta en políticas públicas para su prevención y control. According to IMSS' Family Medicine Information System (SIMF) Institute has record of treating 2.4 million diabetic beneficiaries during 2014.

[5] PrevenIMSS has increased its coverage significantly in the past decade, in 2006 through this program IMSS gave 8.8 million check-ups, by 2014 it gave 28.8 million.

Figure 1
DM-2 prevalence for OECD members

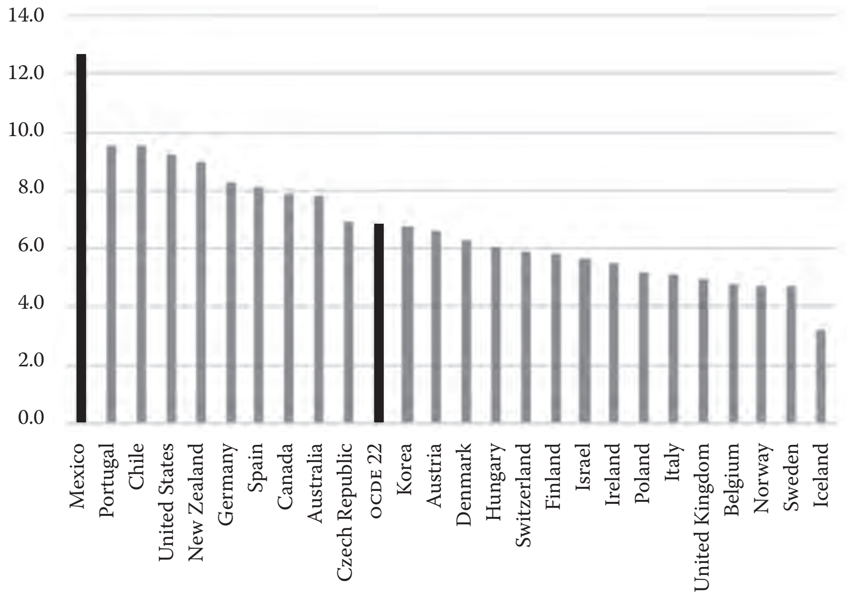
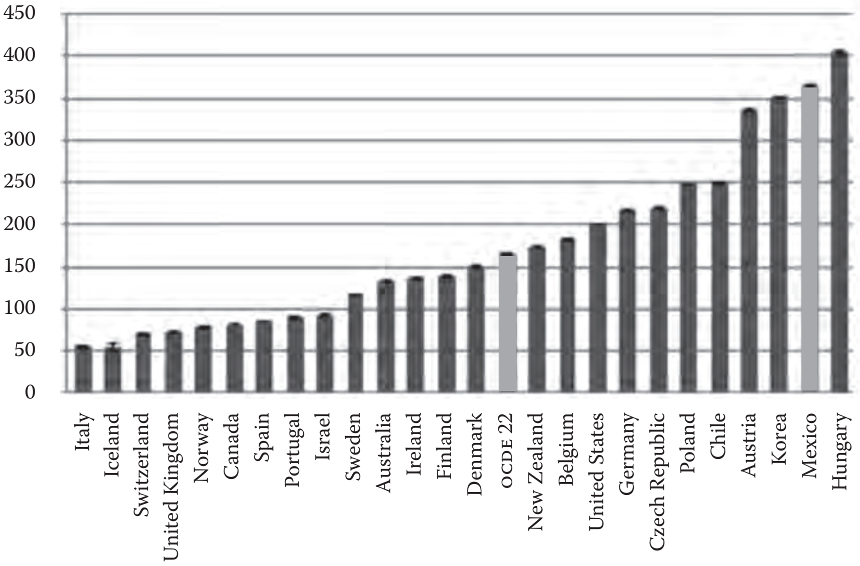


Figure 2
Hospital admissions related to DM-2
per 100,000 people



IMSS is a public social security institution and the country's largest healthcare provider, with a roster of over 450 thousand employees. It is also the largest health-insurer, with over 58 million beneficiaries (current and retired workers and their families as well as insured students). The Institute's services also include childcare at over 1,400 locations nationwide, and management of pension funds for maternity and work-related illnesses, as well as several recreation centers. The institute works directly with the Mexican government's cash-transfer program –the largest conditional economic transfer program in Latin America–Prospera, which offers free access to healthcare to around 12 million people in need.

On a regular day, IMSS personnel attend about 500 thousand medical consultations, perform 4,000 surgeries and deliver 1,200 babies, while over 200 thousand infants are attended at the Institute's childcare services. IMSS is funded by a three-part quota system, with per-capita contributions coming from workers, employers and directly from the federal government. It is the second largest tax-collector in the country –behind only the Ministry of Finance's Tax Administration Service (SAT)–having amassed over 220 billion Mexican pesos (1.3% of GDP) in collections during 2014.

IMSS operates 1,122 first-level medical units (family doctor units, UMFs), 381 auxiliary first-level units, 246 hospitals and 36 high specialty clinics (not including Prospera). In total there are over 1,500 locations where first-level healthcare is available within IMSS' provision net.[6] Between 2012 and 2014, the Institute's wide infrastructure provided first-level medical care to over 35 million individual beneficiaries, together with 6 million attendees to secondary-level specialty clinics.

In 2013, the International Diabetes Federation (IDF) estimated 382 million people with DM-2 worldwide, and predicted that by 2035 this number would increase to 592 million; that is, 8.8% of the world's adult population. The IDF also estimates that 80% of diabetics live on low and middle income countries and argues that rising prevalence is caused by urbanization, aging and lifestyle changes.[7] In our view, Mexico clearly fits the profile of a diabetes-prone region.

In 2010, complications associated with DM-2 were the second most common cause of death in Mexico, and ranked as the top-5th source of disability adjusted life years (DALYs).[8] Based on 2012 data from Ensanut, the National Institute of Public Health (INSP) estimated that 6.4 million adults have been diagnosed with DM-2, representing only half of all individuals estimated to be living with the pathology

[6] 2013-2014 Financial and risk status report to the executive power and congress, Mexican Institute of Social Security.

[7] Guariguata, L., *et al*. "Global estimates of diabetes prevalence for 2013 and projections for 2035." *Diabetes research and clinical practice* 103.2 (2014): 137-149.

[8] Feigin, V. L., *et al.* "Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010) and the GBD Stroke Experts Group. Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010." *Lancet* 383.9913 (2014): 245-54.

in Mexico, resulting in prevalence of 14.4% on Mexico's adult population,[9] which contrasts with 4% in 1993[10] and an estimated 6.4% on the European Union.[11]

Funsalud estimates that in 2013 DM-2 cost the equivalent of 2.25% of Mexico's Gross Domestic Product (GDP), with 1.11% coming from direct costs (medical attention) and 1.14% from indirect costs (premature death, loss of productivity and absenteeism among others). DM-2's cost estimation for 2018 is expected to amount to 2.62% of the country's (GDP), which highlights the importance of attacking DM-2 more actively.[12] On the same study, Funsalud makes some alarming clarifications among direct and indirect costs. On one hand, treating DM-2 complications amounts for 87% of the money spent on medical attention for the disease while premature death amounts to 72.5% of its indirect costs.

Without constant and persistent check-ups, DM-2 is hard to identify at an early stage with the best preventive measures usually exceeding traditional healthcare responsibilities and involving significant habit changes on the part of individual beneficiaries. There is clear consensus that prevention is the best approach to attack the main causes of chronic degenerative diseases, as it results in the best strategy to reduce incidence, prevalence and mortality and thus can subsequently limit expenditure in patient-care. Alike any other public healthcare provider, and as the prevalence of chronic non-communicable diseases like DM-2's has increased substantially over the past few years, IMSS has been in need of adjusting its service platform in order to better cope with non-communicable as opposed to transmittable diseases.

PrevenIMSS, a national program launched in 2002 has been the Institute's most comprehensive tool to raise awareness towards prevention and improve detection.[13] However, within this vast program there is considerable scope for improvement, mainly in terms of effective detection planning as well as on vehicles for stratification which can warrantee channeling patients to their best fit for attention. During 2014, PrevenIMSS managed over 30 million visits, 3 million more than the previous year and about 4 times the amount registered in 2004. A typical preventive check-up consists of a general health evaluation, with patients receiving information about health improvement actions. In order to improve early detection of patients at risk for DM-2, a capillary glucose test is also applied. The current administration at IMSS has already expanded outreach of this program via a major

[9] Villalpando, Salvador, *et al*. "Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population: a probabilistic survey." *Salud pública de México* 52 (2010): S19-S26

[10] Tapia-Conyer, Roberto, Héctor Gallardo-Rincón, and Rodrigo Saucedo-Martinez. "CASALUD: an innovative health-care system to control and prevent non-communicable diseases in Mexico." *Perspectives in public health* (2013): 1757913913511423.

[11] Whiting, David R., *et al.* "IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030." *Diabetes research and clinical practice* 94.3 (2011): 311-321.

[12] Barraza-Lloréns M., Guajardo-Barrón V., Picó J., García R., Hernández C., Mora F., Athié J., Crable E., Urtiz A. (2015) Carga económica de la diabetes mellitus en México, 2013. México, D.F.: Funsalud.

[13] *Manual de indicadores de dotación de fuerza de trabajo.*

media campaign called *Chécate, Mídete, Muevete* (“Check, Measure and Move yourself”), that reminds people that prevention and active lifestyles are the best way to avoid having to live with the pathology.

According to our figures, on December 2014 IMSS diagnosed 3.3 million DM-2 patients, well over Ensanut’s estimation. However, a significant amount of IMSS beneficiaries do not receive preventive check-ups or do not complete the process necessary to be subjected to a confirmatory laboratory diagnostic test. This is the case because the process elapsing between an out-of-range capillary glucose observation (maybe at a PrevenIMSS booth located at the entrance of most IMSS first-level clinics) and a diagnosis confirmation at the lab is significant as it involves considerable waiting times at the medical unit.

Detecting DM-2 patients at an early stage can make a major difference in their life and can have a large economic effect. This is especially the case at IMSS since the Institute is obliged to pay a per-diem incapacity fees to every beneficiary that falls ill and who is unable to work due to complications related with the disease. Understanding the urgency of the situation, the current administration at IMSS has led an unprecedented effort to detect beneficiaries at risk of becoming diagnosed with DM-2. One of the major goals of this initiative has been to prevent the disease from spreading further via an active focus to reduce hospitalizations through early stage detection. The current study attempts to offer evidence that in order to provide cost-effective solutions to this sort of challenge the Institute’s datasets and econometrical analysis can be a key element of the strategy.

In what follows we describe the process by which the Institute’s data bases can be utilized to detect risk-prone patients. Emphasis is made on how IMSS structure and unique mix of individual data on health and socioeconomic characteristics can be exploited in order to predict risk of DM-2 diagnosis without performing additional medical tests. In particular, we go in detail about how we combined a database of over 50 million observations containing socio-economic variables, with another information-set populated with each and every diagnosis made at an IMSS family clinic between 2012 and 2014. This process involved the amalgamation of 245 million registries together with a structural database that characterizes the 1,229 established first-level clinics.

In the current study we dwell also on possible implementation policies of our results and make an attempt at estimating the impact that an outreach strategy to detect potential diabetics could have on IMSS’s capabilities to handle the disease effectively. This is a key part of the paper as it shows how the institute’s centralized Big-Data information organization structure is capable of leading towards cost-effective easy to implement measures that can improve resource allocation and maximize the impact on healthcare.

The rest of the paper is organized as follows. We first describe related literature on the subject, to then go into explaining the details of the data we used in order to construct the main variables that play a role in our risk assessment model. Next we explain how we adjusted the model to distill the results obtained as well as their estimated effect on imss's dm-2 detection capabilities. As the current document is by no means a concluded piece, we finish with some suggestions for further research that would allow for the application of the methodology to other chronic-degenerative diseases.

Literature review

According to our databases, the number of imss adult beneficiaries diagnosed with dm-2 reached 3.3 million by December 2014. With an increasing prevalence, the current administration has felt the imperative need to enhance the Institute's detection ability thus directing an unprecedented effort to exploit the datasets that would help device an effective solution to handle the spread of the disease. The results presented in this document constitute in our view a definitive contribution to that objective, as they provide an econometric tool capable of identifying dm-2-prone individuals early on in the game. On the one hand, these techniques propose a mechanism that can be used to classify by risk category all imss beneficiaries at a given point of time. To perform such a feat we depart from a subset of those beneficiaries whose rights were still current on May 2014, and that attended a family clinic between 2012 and December 2014. Based on our results we were also able to define a simple questionnaire, consisting of 4 data points, which allows for calculating the risk of dm-2 diagnosis for those individuals outside of our sample, thus allowing for recently incorporated imss beneficiaries or unregistered visitors. This can be done even based on the model's probability estimations by leveraging the Institute's data configuration, and without the need of any additional medical tests.

The present study utilizes a binary response generalized linear model (glm) to estimate the probability of becoming diagnosed with Diabetes Mellitus Type 2 (dm-2) for the population of beneficiaries of Mexico's Social Security Institute. In order to reach such predictions, we use data from the Integral Family Medicine System (simf) at imss, which gathers information from the Institute's 1,229 simf-connected first-level medical units distributed across the country. Specifically, we adjust a logit econometric model that predicts such risk for a population of 17.5 million adults that acquired at least once the services of a first-level family doctor between the years 2012 and 2014. It must be noted that the subsequent analysis is restricted to the sub-set of imss' beneficiaries registered within the system in May 2014. In what follows, we not only present the results of our estimation, but also discuss the stability of the model's estimates, and how these can be used to predict risk for individuals that did not attend such clinics during the period analyzed. The latter implication would allow us to widen the outreach of our risk-profiling to a population of over 58 million.

glm analysis is still an active research area employed repeatedly on recent studies. Hosmer and Lemeshow (2004) provide a detailed explanation of the model utilized in this document and the conditions under which it can be applied empirically. Cepeda, Boston, Farrar and Strom (2003) discuss the advantages of this logistic approach –as an improvement over a propensity score methodology– whenever the number of events is more than 8 times the number of confounders, a condition that fits well within the features of our datasets. Under the context of the methodology for econometric estimation utilized here, and using a Montecarlo approach, Bergtold, Yeager and Featherstone (2011) perform robustness checks for different sample sizes and find that samples of over 250 observations significantly reduce estimation bias.

On the present investigation we estimate models with over 10.5 million observations for each of the years considered, which according to previous studies implies scant estimation bias. The contributions of our analysis to the empirical literature on the subject are evident as we show that a large prediction model can produce statistically stable estimators even in the presence of dynamic fluctuations. Our results prove that a GLM econometric approach can be utilized for identifying chronic disease risk-prone individuals, and can produce a simple and well behaved risk prediction model. With a handle on these outcomes, we leverage this methodology to suggest cost-effective policies potentially able to improve preventive medical attention at IMSS clinics.

There have been many empirical studies utilizing GLM based models to predict risky events: Kimball, and Dietriech (1984) use a logit model to predict the probability of mergers, and manage to classify 92.4% of the firms correctly. Martin (1977) uses a logistic regression approach to predict bank failures. Using topographical factors, Lee (2005) employs a logistic regression model to estimate the risk of landslide at Penang, Malaysia. Valenzuela, Roe, Cretin, Spatie and Larzen (1997) also make use of logistic regression to predict the survival of intervention from cardiac arrest. These are only a few examples of research that contributes to the broad literature that applies GLM models to obtain inferences on risky shocks.

On research closer to the topic at hand, Narayan, Boyle, Thompson, Sorensen and Williamson (2003) estimate the risk of contracting DM-2 during a person's lifetime using race, age and sex as key factors, and then apply their results to arrive at an estimate of the cost of the disease. Tapia, Gallardo and Saucedo (2013) discuss an initiative proposed by the Slim foundation –a Mexican private entity– so called Casalud. Based on a systematic three-step risk assessment, in which a questionnaire that utilizes the body mass index (BMI), age, sex, blood pressure and waist circumference is used to detect obesity and the likelihood of hypertension. In the particular cases where an individual presents 5 or more risk factors, the study incorporates a measurement of capillary glucose in order to identify the person as pre-diabetic (>100mg/dl) or diabetic (>125mg/dl). A third-stage test uses a measurement of serum creatinine to estimate the likelihood of kidney disease.

The study that is most related to ours is an effort by Akter, Misanur, Rahman, Krull and Sultana (2014) from the World Health Organization (WHO), which tries to identify, also via a logistic econometric model, the factors that affect the probability of being DM-2 positive for the Bangladeshi population. This study utilizes a representative survey that includes socioeconomic, community and health variables for a group of almost 8,000 individuals. The analysis we present here improves over the latter investigation as it makes use of the whole population with nominative datasets, hence being able to predict the probability of being DM-2 positive for each individual. Moreover, our study also incorporates genetic-heritage and medical infrastructure variables. Finally, the analysis lends itself to show that the econometric model estimated is statistically stable over time as its predictive power encompasses a population from 2 years back. This latter feature suggests that our methodology can be used to assess the risk of individuals beyond the dataset.

Summarizing, to the best of our knowledge the present document provides three main contributions to the DM-2 literature. First, it is an unprecedented effort to incorporate structural variables, like accessibility of health attention provided at the medical unit that can thus be thought of as indicators of demand pressure for services. This is relevant since the probability to develop a certain disease (or to be provided with diagnosis) is likely to be affected by the system's structural capacity to offer patient care. Second, the findings hereby summarized constitute an initial attempt to propose a tool for stratification of risky patients that utilizes a large nominal dataset consisting of socioeconomic and medical history variables. This latter effort becomes even more relevant as we extend the outcomes from the econometric model to predict the risk of DM-2 on individuals outside the sample without the need to perform any additional tests. Finally, it helps to illustrate that a predictive model can remain statistically stable through several years of estimations performed independently on a large population.

The data

For this paper we use a nominal dataset that combines medical/health, family composition, structural and socio-economic related variables managed by different areas within IMSS. Our study uses a full nominal dataset (not a sample) to approximate the probability of being diagnosed with a disease, in our case DM-2. In this section we describe our main datasets and the processes by which we filtered and merged each component. We conclude this section by describing the final datasets with which we proceed to implement the subsequent econometric analysis.

Datasets

Medical health characteristics

The Primary Health Attention Unit (UAPS) is in charge of all first-level medical services at IMSS. UAPS provided data on each and every diagnosis made at a first-level clinic between 2012 and 2014, as well all the capillary glucose test results detected at the Institute's prevention wing, PrevenIMSS. The data amounts to 245 million diagnosis and 30 million capillary glucose measurements undertaken on 35 million individual beneficiaries during the years 2012-14. We were thus able to gather basic medical characteristics for each and every individual that attended a first-level medical clinic connected through the data SIMF (Family Medicine Integral System, Sistema Integral de Medicina Familiar) during this time-span. These main medical characteristics include: weight, height, age, gender, blood-pressure, glucose concentration in blood, and every diagnosis arrived at by a first-level physician at IMSS.

IMSS diagnosis are reported based on the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), which facilitate the handling of diagnosed diseases data at the individual level. Through the richness of this data we were able to identify risk factors such as hypertension, elevated levels of cholesterol and triglycerides. Height and weight measures allow for the calculation of the body mass index helping identify overweight and obese individuals using World Health Organization standards. To create a proxy for cultural eating habits, we accounted for the number of overweight family members. In order to generate a medical history indicator (that would capture the hereditary content carried by each individual) we took advantage of the data structure at IMSS, which allowed for the identification of each beneficiary's family members diagnosed with a disease at any clinic within the system during the previous three years.

Medical supply characteristics

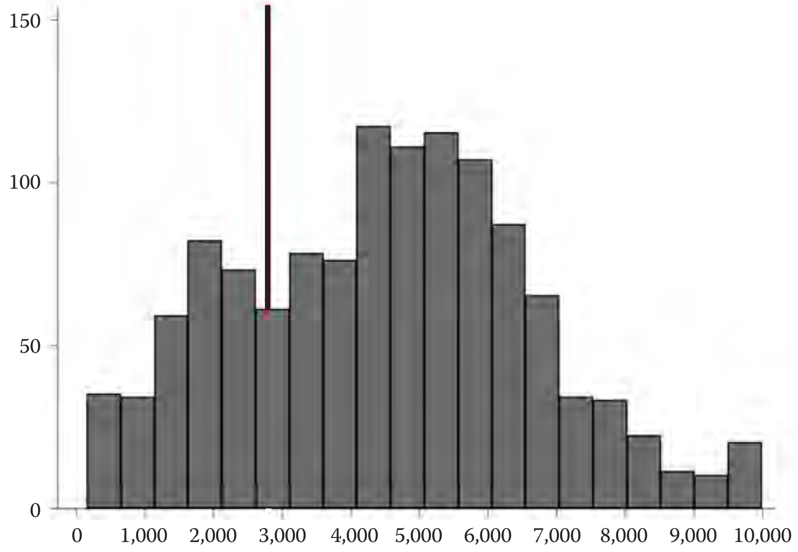
IMSS has a diverse net of medical facilities, ranging from small clinics within rural communities to some of the largest specialty hospitals in the country. The present study utilizes data from the 1,229 first-level healthcare facilities at IMSS to estimate the probability of having a positive DM-2 diagnosis. In Mexico, access to private medical facilities is limited as it is directly linked to income. Given that waiting times associated with access to public healthcare are on average substantially longer than those at private clinics, in many cases the inability to access preventive medical attention can directly affect the likelihood of disease.

In order to incorporate this dimension into the estimation, we account for the number of beneficiaries[14] subscribed to family clinic at each medical unit and divide it by the number of doctor's offices within the unit. IMSS guidelines[15] recommend that a medical unit (UMF) does not exceed 2,800 members per doctor's office. However, we find that 983 out of the 1,229 exceed this benchmark, with some units even reaching over 18,000 patients. In figure 3, a histogram for this variable is displayed, for which the vertical axis captures a measure of the frequency (number

[14] Rights-holders at IMSS are all formal workers registered by their employees at the Institute and their direct families. Employees and workers contribute each month with a fraction of their salaries to the Institutes' revenue.

[15] *Manual de indicadores de dotación de fuerza de trabajo.*

FIGURE 3
Subscribed beneficiaries per medical office



of medical units), with the number of beneficiaries subscribed per medical unit accounted on the horizontal axis, with the red vertical line pointing to the 2,800 benchmark recommended by the Institute's guidelines. Finally, all medical units with over 10,000 beneficiaries were accumulated on the last bin (13):

Time of travel between a medical unit and its associated hospital (usually located in large cities) is utilized as an additional structural variable to capture access to medical services. Estimated time ranges from 0 minutes (whenever the first-level medical units are a part of the hospitals or are located within the same facility) to over one day. Figure 4 contains a histogram showing the frequency of medical units relative to the time it takes to travel from them to their associated hospital unit. All medical units with over 200 minutes of traveling time are included in the last bin (47 in total).

Socio-economic, work related and family characteristics were obtained from the Incorporation and Revenue Directorate (DIR, Dirección de Incorporación y Recaudación). Individual information was provided for each IMSS beneficiary as of May, 2014. The database contains 51 million individuals, over 40% of the country's population, and includes characteristics such as: daily wage, working sector and subsectors,[16] years of experience, gender, role in the worker's family together with the medical unit to which the worker is subscribed (depending on location).

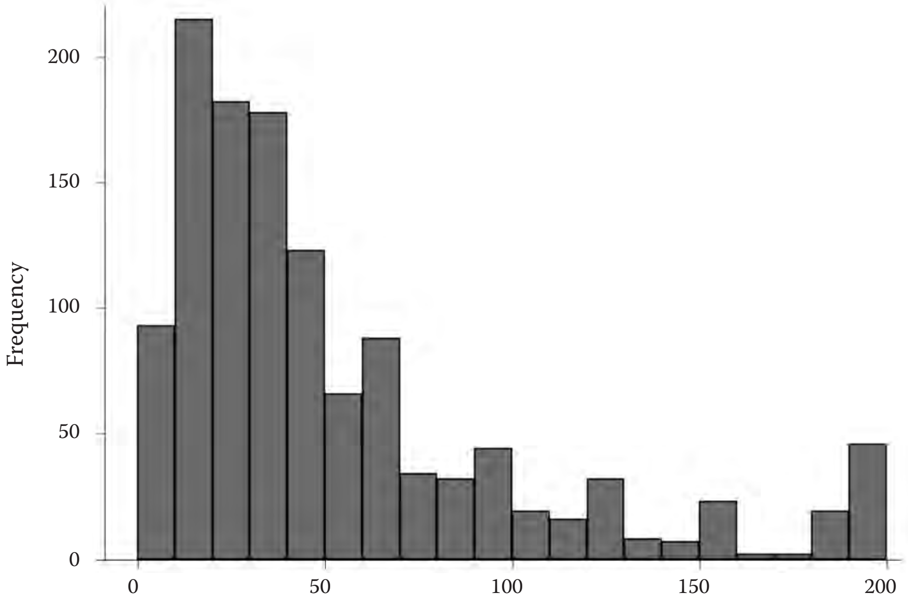
IMSS provides health and work-related injuries insurance to workers and their economic dependents with a direct family link.[17] Workers are registered into the system by their employer, with each worker assigned a social security number

[16] IMSS considers 9 sector and 99 subsector classifications for its workers.

[17] Economic dependents are limited to spouse, children and parents.

(NSS), and with every family member he/she introduces into the system included into the DIR database with a Medical Aggregate number identifying his/her role within the family. By exploiting this construction we are able to identify family size, the income of the family-head together with all medical indicator variables previously discussed.

Figure 4
Time of travel from medical unit to hospital of reference



Data merging and cleansing process

In this section we describe the merging process undertaken in order to build a database that could be utilized for estimation. Given that the datasets described so far were provided by different so-called jurisdictions within IMSS, it turns out that the effort to exploit these amalgamated datasets is unprecedented.

We handle first the medical health datasets in order to construct diabetes and hypertension disease dummies for each individual (hypertension is a well-known risk factor for diabetes). We do this by identifying whether an individual beneficiary received at least one diagnosis for any of these two diseases.[18] We thus search within 80 million observations accountable for each of the years considered.

The body mass index (BMI) variable is generated utilizing an individual's height and weight and combining them according to the formula: BMI = $\frac{weight}{height^2}$. It is relevant to note that whenever the BMI measure refers to an adult, observations

[18] IMSS databases containing diagnostics are coded based on the International Statistical Classification of Diseases and Related Health Problems, 10th revision.

under 15 or over 100 are considered as likely typos. However, being overweight is an important risk factor for diabetes, and we therefore made an additional effort to lose as few observations as possible. Whenever presented with a BMI value under 15 or over 100, we utilized the mean of any other BMI measures for that same individual and kept the most recent observation.

After dropping observations with unfeasible BMI data, and having one observation per person, we generated family related diagnosis variables for DM-2. We also created a variable indicating how many overweight and obese individuals each beneficiary is associated with within his/her direct family. As DM-2 rarely manifests itself previous to adulthood, we dropped from the estimation all individuals under 18 years of age.

Drawing on data from IMSS's prevention wing, a capillary glucose measurement was incorporated into the analysis. To avoid a probable change in characteristics and complying with expert advice from doctors, we incorporated measurements only with the diagnosis datasets from the same year they were taken and, kept the mean of the scores whenever there was more than one measurement. We thus obtained indicators of capillary glucose levels for 4, 4.2 and 4 million DM-2 diagnosed individuals for the years 2012, 2013 and 2014, respectively. Unpaired observations were not dropped since having a preventive checkup is not mandatory; in order to avoid affecting the glucose coefficients by the fact of being tested, we control instead by having a preventive check-up measurement as well.

Drawing from the socio-economic, work and family dataset, family income and family size variables were constructed by exploiting the family identification variables, which was made possible given the way the socio-economic data is structured. We kept variables such as experience, type of insurance (regular, student or pensioned), the sector in which the individual works, morning or afternoon service to which she/he is subscribed and role within the family, with the last one being a categorical indicator containing assigned to head of the family, partner, mother, father, daughter or son. Since we only have the May 2014 work related and family dataset for IMSS's beneficiaries, we utilized it for every year on the regressions. We do not consider this to be an issue since for most IMSS registered workers salary does not fluctuate significantly within the realm of a couple of years.

Taking the three medical/health datasets and using social security and medical aggregate numbers as linking variables, we therefore merged them with socio-economic, work and family indicators. Finally, the structural medical supply data was incorporated. In table 1 we present the results from this initial merging process.

Table 1
Merge results between work and medical variables*

Year	Medical health	Work and family	Merge
2012	14.2 million right holders	51.3 million IMSS right holders in May 2014	10.6 million right holders
2013	14.4 million right holders		11.6 million right holders
2014	14.4 million right holders		12.3 million right holders

* On this table we show how many observations we have on each dataset and report the number of paired individuals on the last column, for each year.

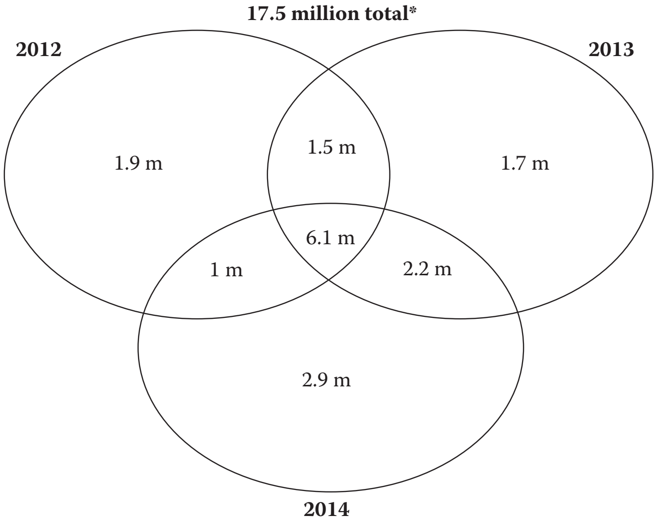
The observations that we lost during this initial merging exercise were largely due to people that lost their rights to IMSS's services. It is highly likely that these individuals migrated outside of our datasets (and were thus employed within the informal sector), particularly since the number of beneficiaries for which we have medical/health data are about the same for all years, and the number of individuals with a full set of variables is smaller as we move further back in time. In order to count on as complete a dataset as possible, we made the choice to consider everyone counting with medical and socioeconomic data, mainly since we want to be able to encompass as many factors as possible into the final estimation. Having been careful during this process proved fruitful when we performed robustness tests. During this process we analyze whether the 2014 model could be usable as a base for inference in previous years. As we were able to predict accurately when applied to 2012 and 2013 databases, we became confident that our cleansing data procedure was rigorous enough.

Finally, the main dataset was completed by adding structural variables. For this last merge, an entity called the spending code (within IMSS' classification) is used as linking variable. This indicator is obtained from the medical unit at which each individual beneficiary is subscribed to. When running this process we lost less than 0.1% of observations for each year, mainly due to mistyping or missing the medical unit's spending code.

Figure 5 illustrates how IMSS' beneficiary population interacts with first-level medical services data from year to year. The Venn diagrams included in the figure display the population that attended family medicine doctor's appointments and received a diagnostic within our database between 2012 and 2014. The 6.1 million right holders in the central area (35%) received first care that resulted in a diagnostic, on all 3 years. Additionally, we can see that 6.5 million (37%) had only one diagnosis in three years, and 4.7 million (28%) had two diagnosis in these three years.

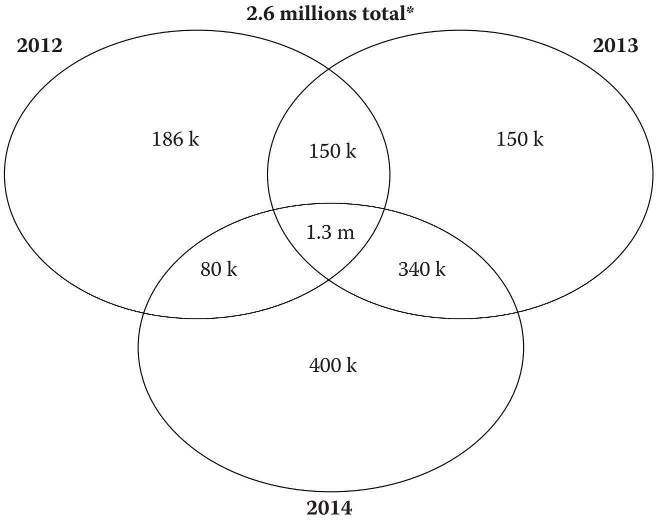
Figure 6 illustrates the first level medical services attendance for DM-2 diagnosed beneficiaries for these same three years. We can see that 1.3 million diagnosed diabetics (50% of the total population within the Venn diagram) attended a first-level medical unit in all three years, with 28% attending only during one year (from which 2014 amounts to only half) and 22% on only two years.

Figure 5
Right holder population attendance to first level medical services between 2012-2014



m = millions
* We only consider adult patients on our final databases.

Figure 6
DM-2 diagnosed right holders attendance of first level medical services between 2012-2014



m = millions k = thousands
* We only consider adult diagnosed patients on our final databases.

The analysis contained in these simple diagrams highlights the enormous variability in the usage of IMSS medical services within and between years, underlining the relevance of evaluating the econometric model's performance each year separately.

Descriptive statistics

In this section we present descriptive statistics for the final datasets that were consolidated for each year in order to arrive at our estimated probability of being diagnosed with DM-2.

First, on table 2 we present mean and standard deviation, when applicable in parenthesis, for age, BMI, percentage of hypertensive patients, percentage of patients with a diabetic relative, capillary glucose measurement, percentage of males, head of family's salary, family size, time of travel between medical unit and it's hospital and population with rights per medical office. All statistics are presented for both patients diagnosed with diabetes mellitus and the rest of the population.

TABLE 2
Summary of relevant variables*

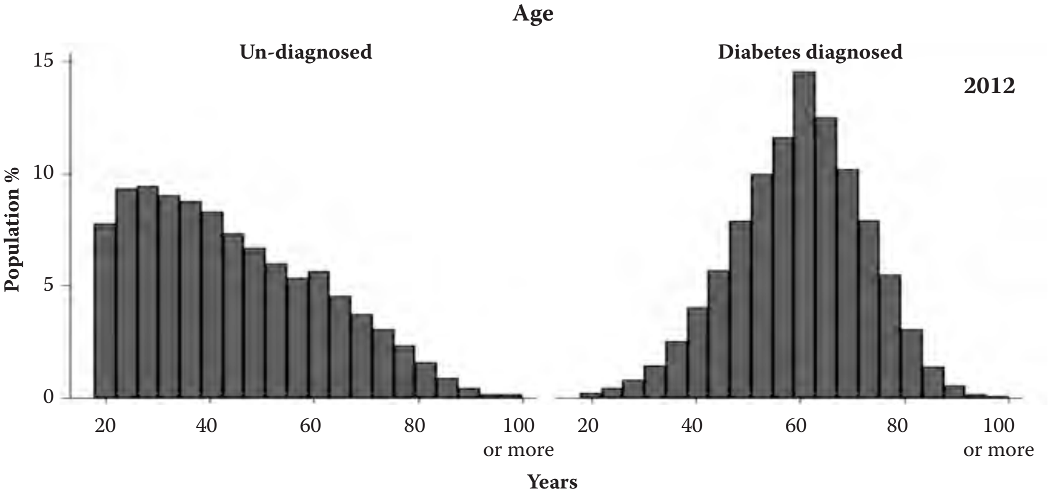
Year	2012		2013		2014	
	Non Diagnosed	Diagnosed diabetics	Non Diagnosed	Diagnosed diabetics	Non Diagnosed	Diagnosed diabetics
Population	8,834,327	1,716,064	9,633,061	1,942,121	10,188,998	2,121,078
(%)	(83.73%)	(16.27%)	(83.22%)	(16.78%)	(82.77 %)	(17.23 %)
Age	44.26	59.95	44.32	60.13	44.50	60.35
	(17.36)	(12.66)	(17.52)	(12.67)	(17.64)	(12.77)
BMI	28.21	29.8	28.19	29.75	28.2	29.75
	(5.38)	(5.53)	(5.41)	(5.54)	(5.42)	(5.58)
% Hypertensive	19.9%	36.2%	20.2%	35.8%	20.6%	35.3%
% Diabetic relative	9%	21.5%	9%	21.2%	9%	20.6%
Capillary glucose	101.67	141.82	101.57	135.61	101.06	131.37
			(33.94)	(71.31)	(33.32)	(68.75)
% Males	38.2%	38.8%	38.3%	39.3%	38.6	39.7%
Family head's salary	290.92	304.67	279.4	296.23	268.54	288.9
	(272.81)	(292.67)	(264.4)	(286.33)	(257.08)	(281.13)
Family size	2.82	2.61	2.76	2.61	2.69	2.59
	(1.5)	(1.41)	(1.49)	(1.41)	(1.47)	(1.39)
Time of travel MU-hospital	178.19	138.152	164.23	124.75	165.6	127.84
	(2,459.54)	(2,105.11)	(2,354.44)	(1,981.1)	(2,366.88)	(2,013.61)
Population subscribed per medical office	5,974.11	5,903.74	5,973.41	5,910.63	5,969.71	5,915.33
	(1,695.4)	(1,629.01)	(1,724.01)	(1,652.36)	(1,726.74)	(1,660.76)

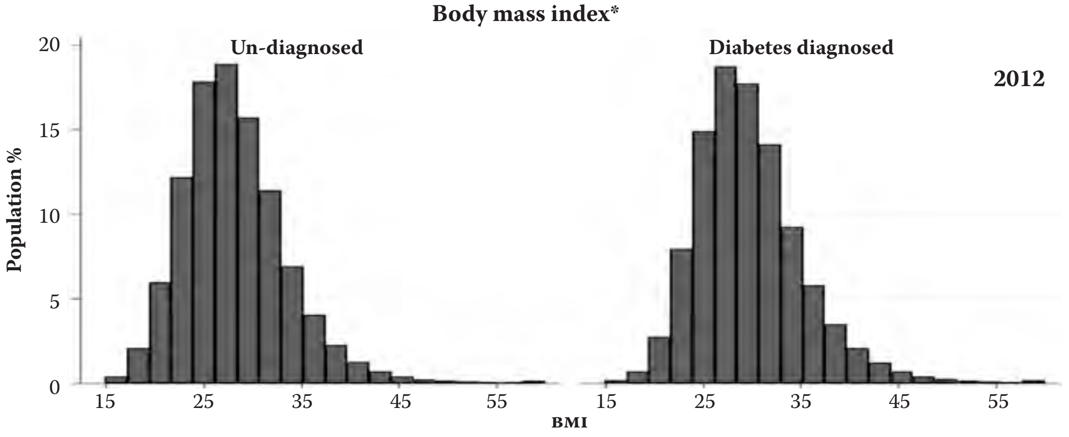
* For the creation of this table we utilize the population from our final datasets for each year individually. That is, we consider population that received a diagnosis on a first-level medical unit and is on our May 2014 DIR dataset for each year.

As can be noticed, the characteristics for the DM-2 *diagnosed* and *non-diagnosed* populations do not have, on average, significant differences from one year to another. From the medical health database we observe that on average the diagnosed population is 15 years older than the rest, and marginally more overweight, much more prone to having hypertension or a DM-2 diagnosed relative and display more elevated capillary glucose measurements. Also, from the work related and family size variables we can see that on average DM-2 patients earn slightly higher income and have marginally smaller families. Finally, we see that, on average, DM-2 diagnosed patients are subscribed to slightly less populated medical units (measured by population per medical office) which are closer to their network hospital.

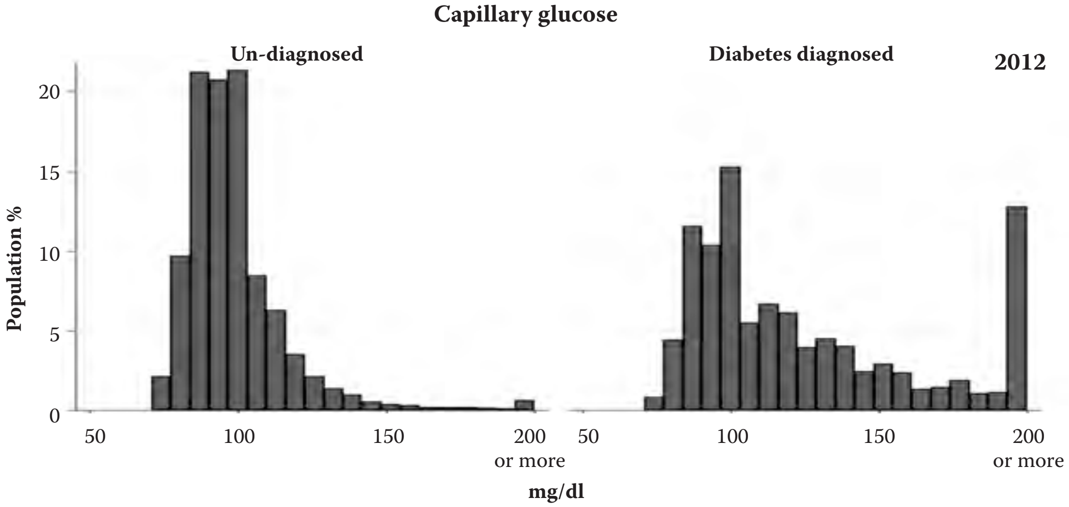
Since some characteristics amongst individuals have a very high standard deviation, we show histograms for certain variables that can aid us in better understanding the difference between both populations. On figures 7, 8 and 9 we present histograms for age, BMI, capillary glucose, head of family's salary, family size, time of travel and population subscribed per medical office for our 2012, 2013 and 2014 population respectively. On these histograms we show percent of plotted population on the vertical axis and the value of the variable of interest on the horizontal axis.

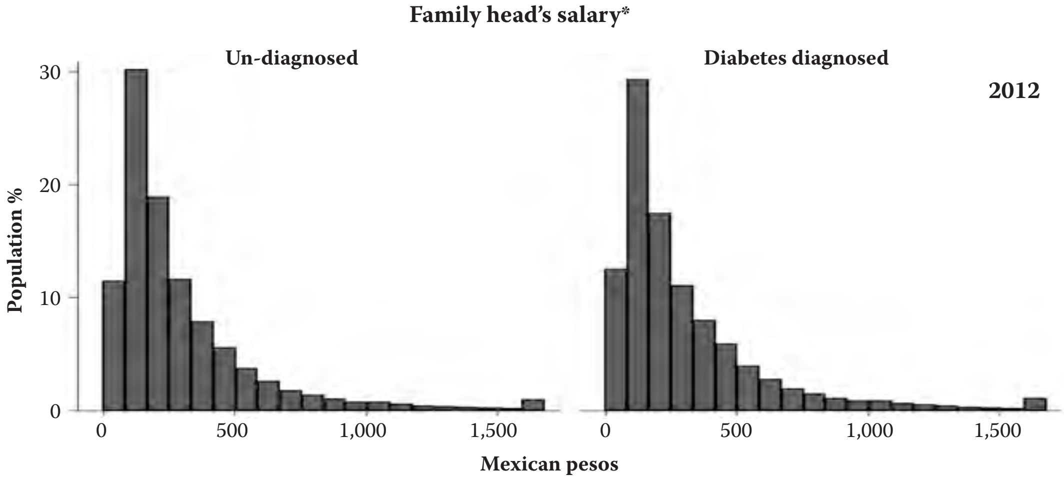
Figure 7
Histograms for 2012



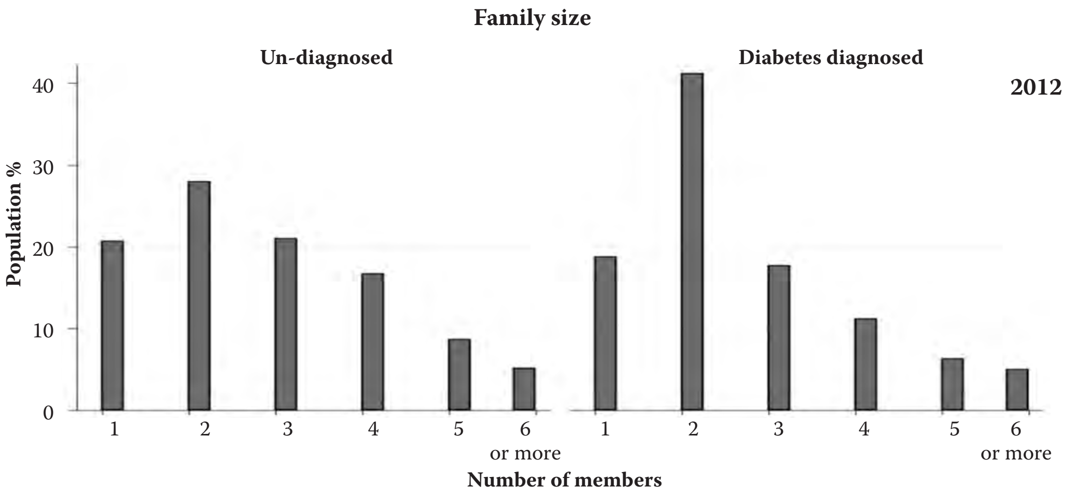


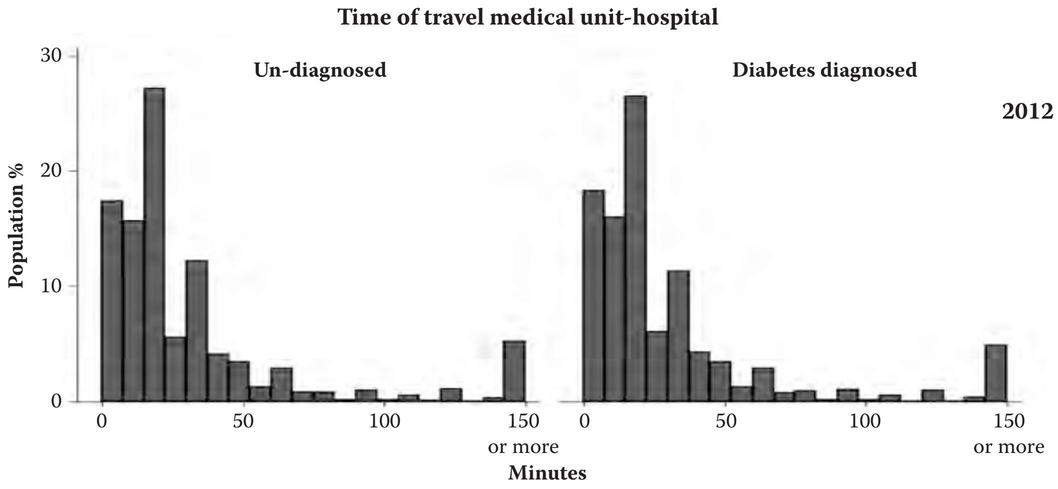
* Any observation over 60 BMI is shown on the last bin.





* Any observation of over 1,682.25 Mexican pesos (25 minimum wages) is shown on the last bin.


Family size
Un-diagnosed
Diabetes diagnosed
2012
Population %
0
10
20
30
40
1
2
3
4
5
6 or more
Number of members


Time of travel medical unit-hospital
Un-diagnosed
Diabetes diagnosed
2012
Population %
0
10
20
30
0
50
100
150 or more
Minutes

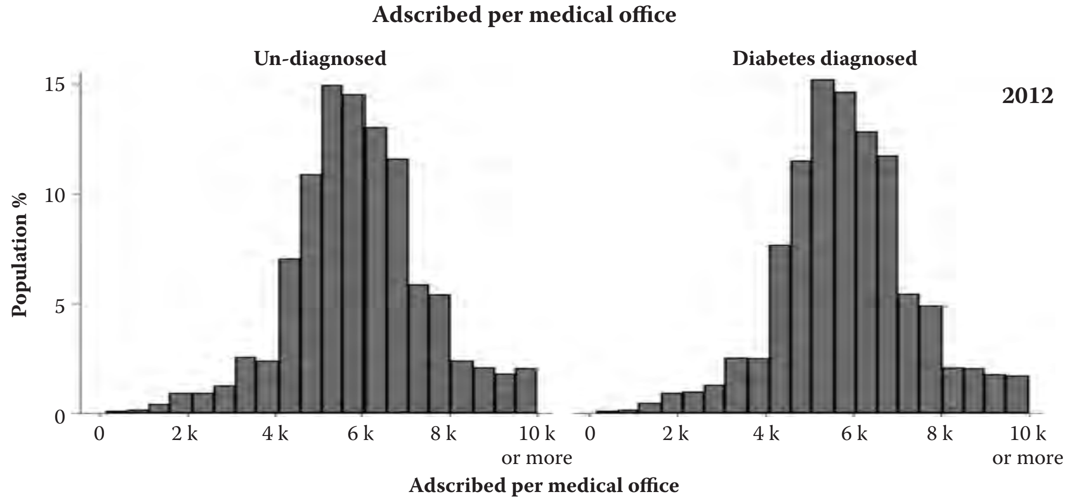
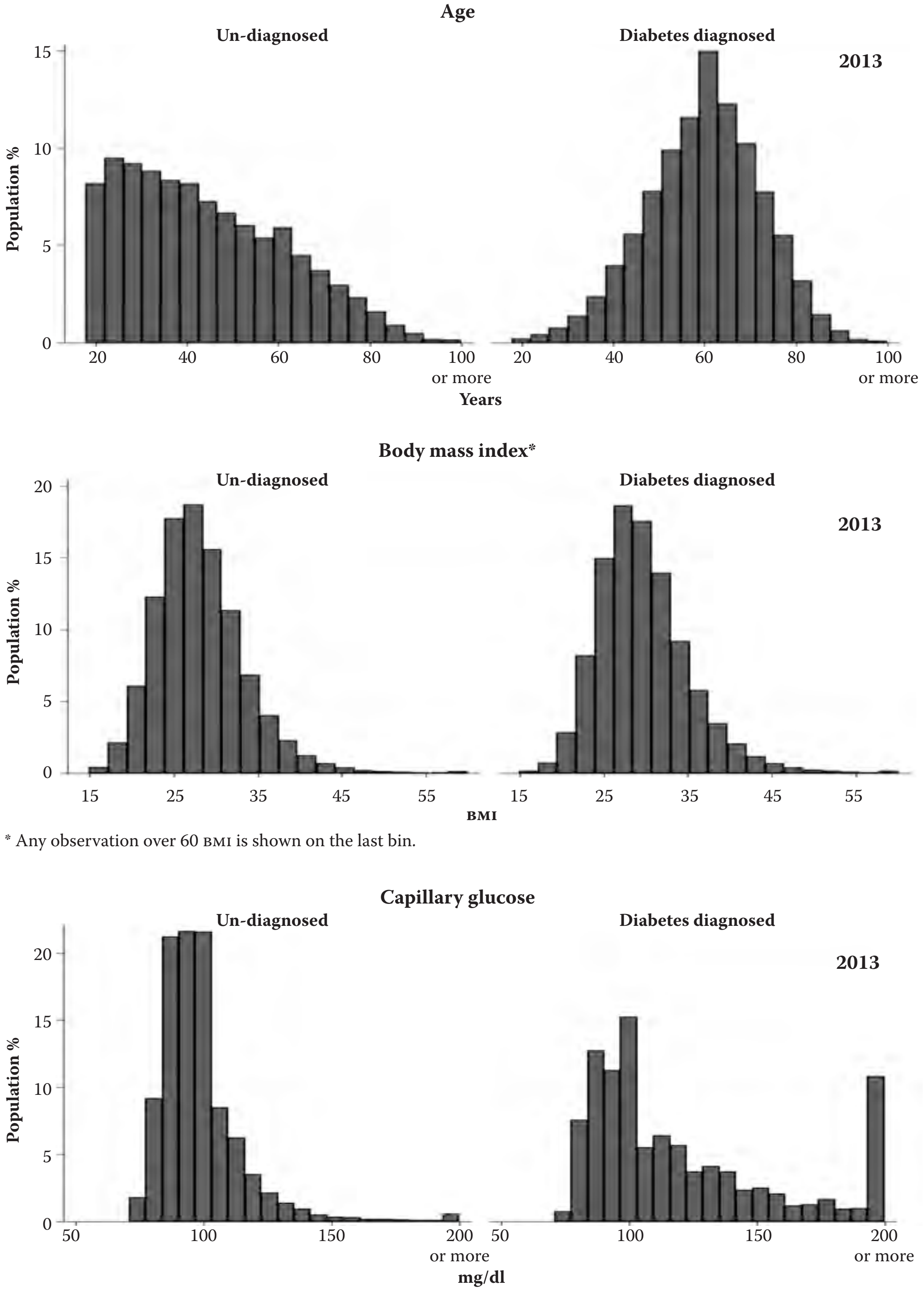

Adscribed per medical office
Un-diagnosed
Diabetes diagnosed
2012
Population %
0
5
10
15
0
2 k
4 k
6 k
8 k
10 k or more
Adscribed per medical office

Figure 8
Histograms for 2013

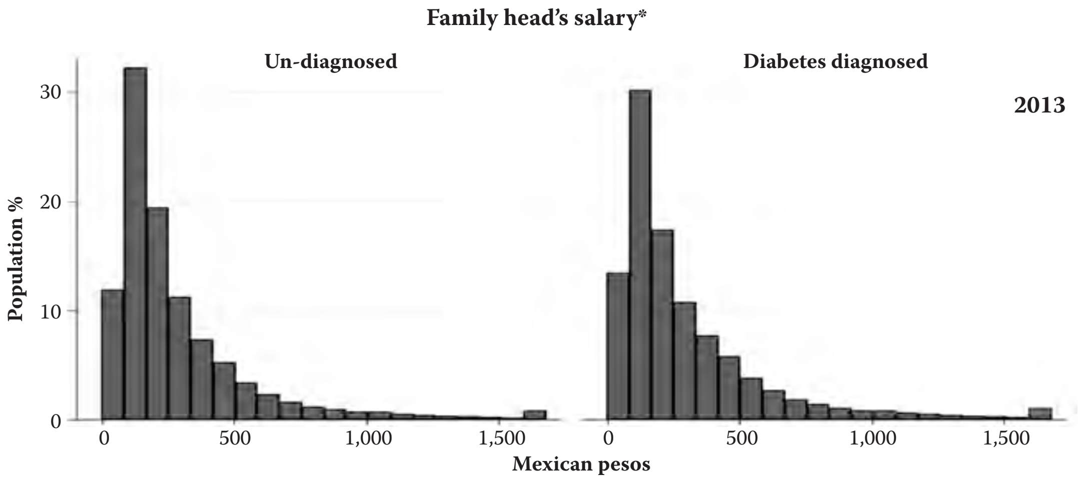


Age

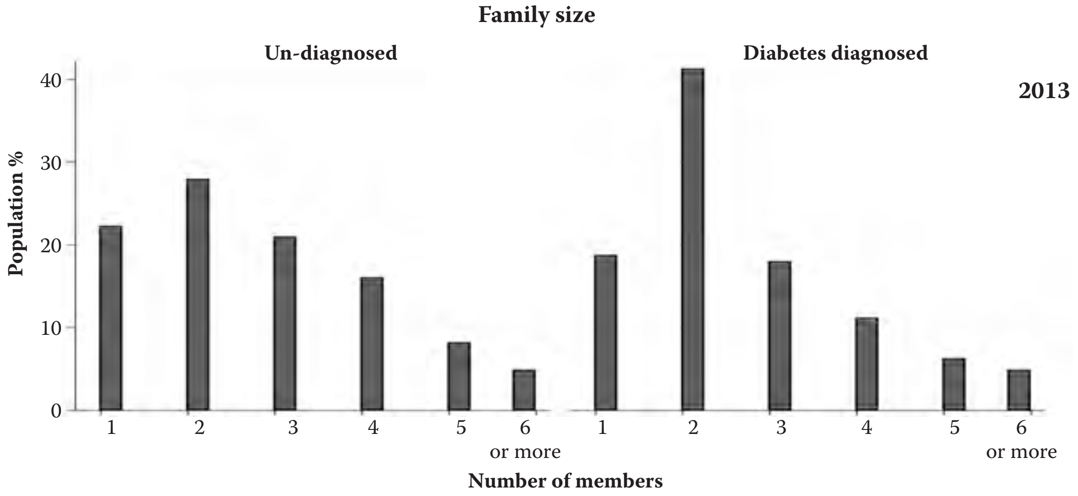
Body mass index*

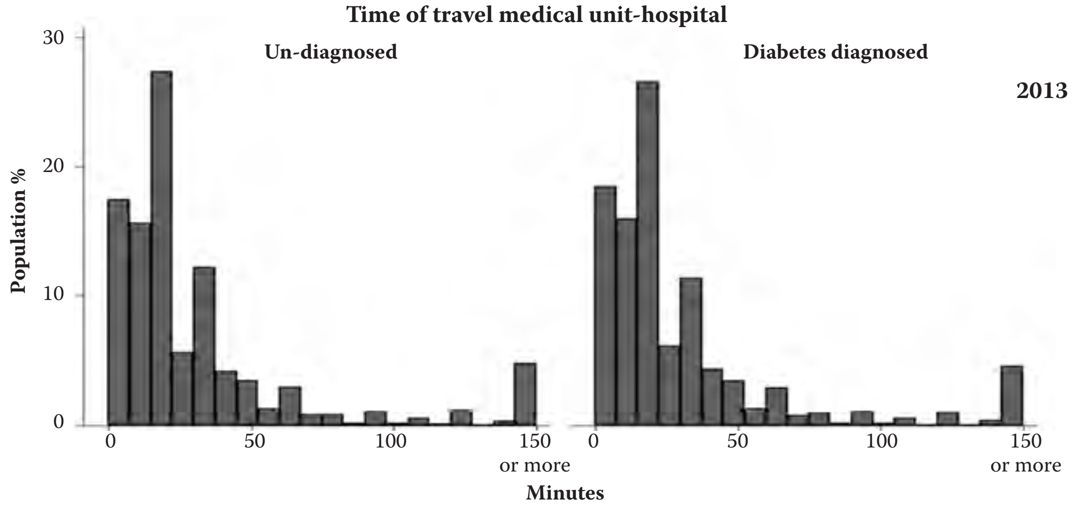
Capillary glucose

* Any observation over 60 BMI is shown on the last bin.



* Any observation of over 1,682.25 Mexican pesos (25 minimum wages) is shown on the last bin.





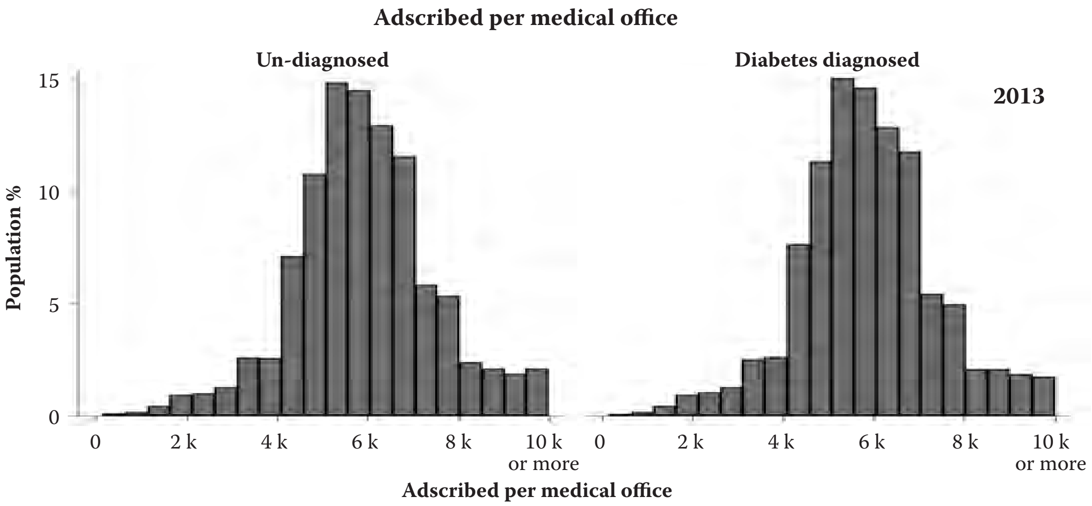
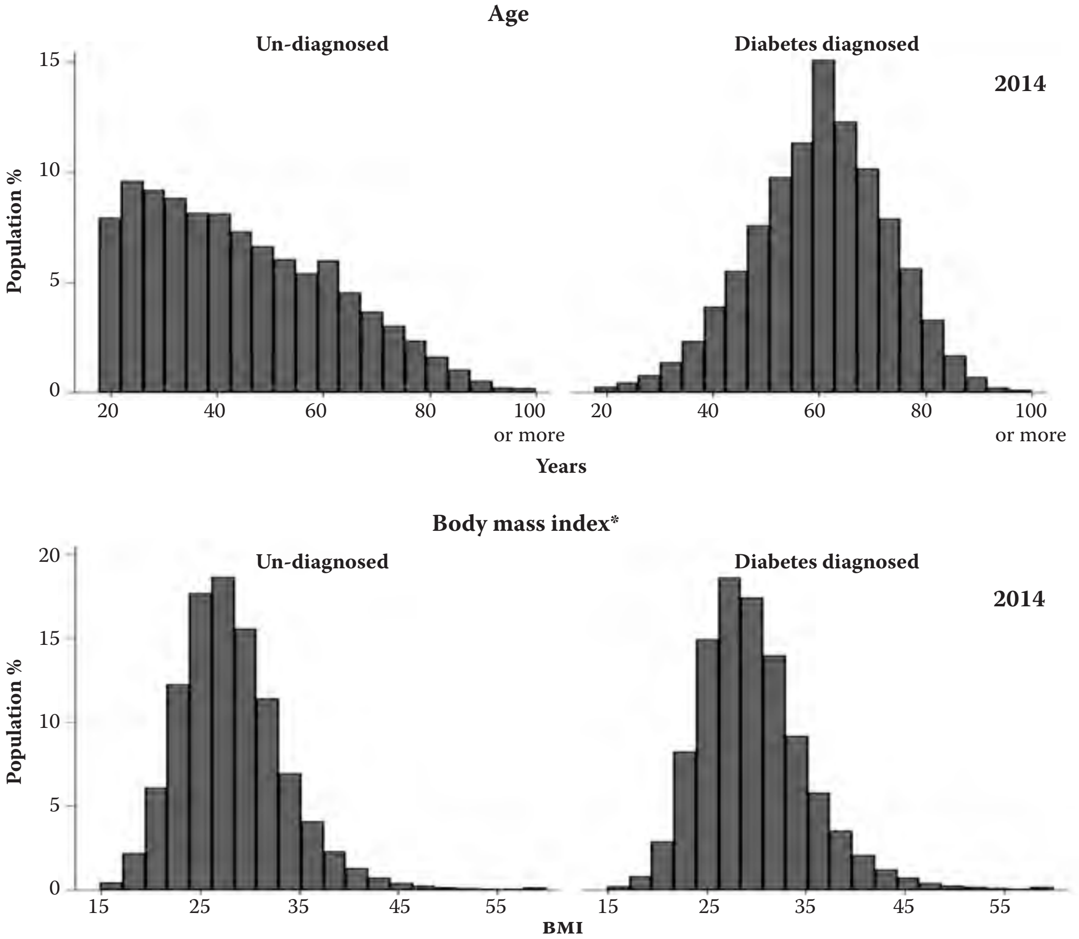
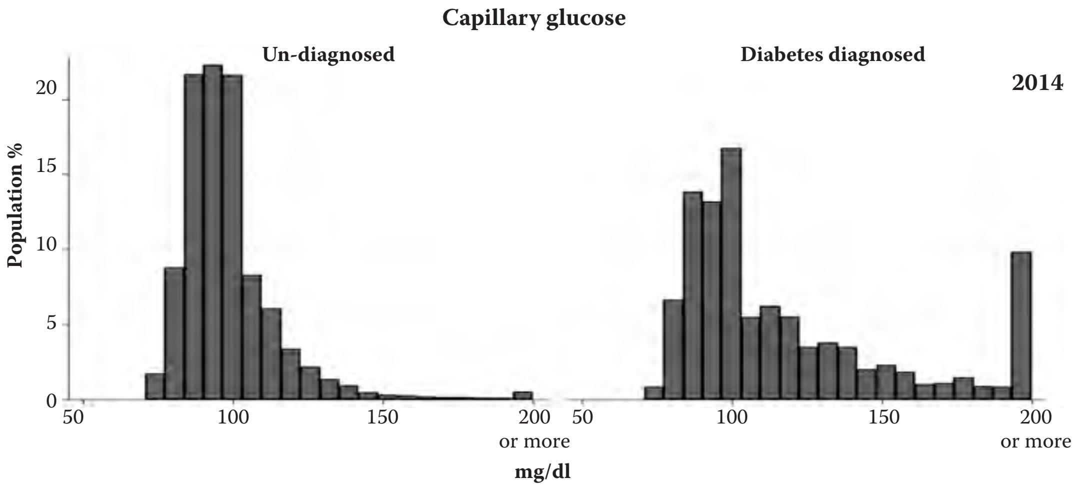
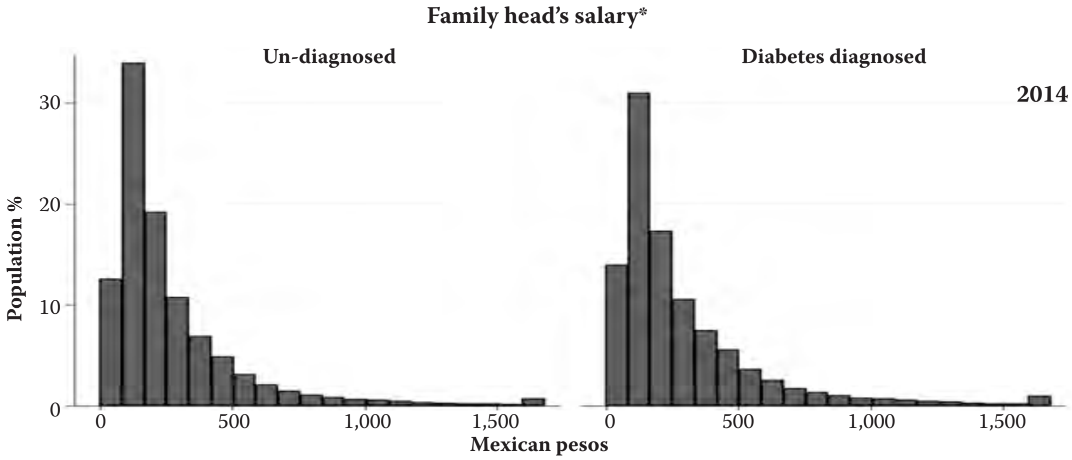


Figure 9
Histograms for 2014

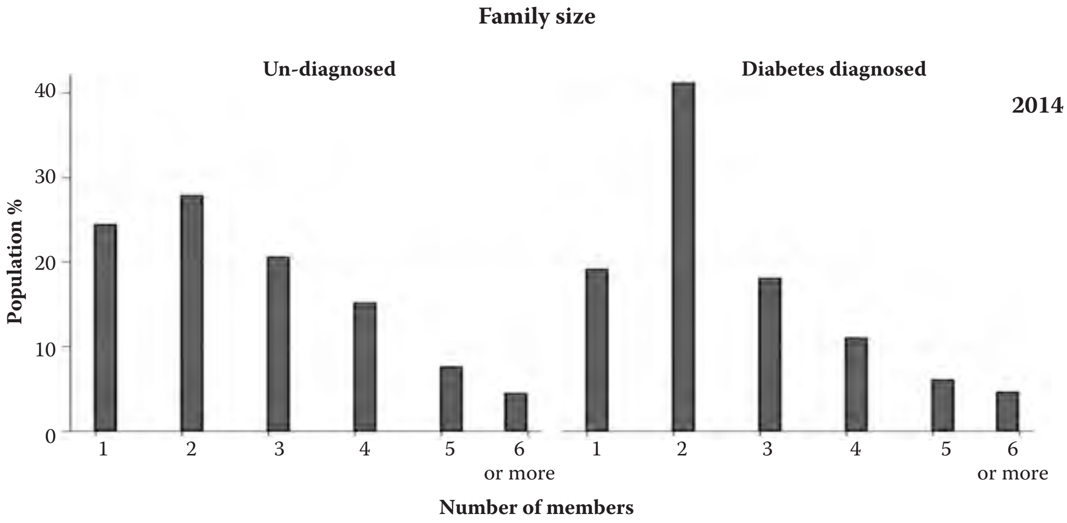


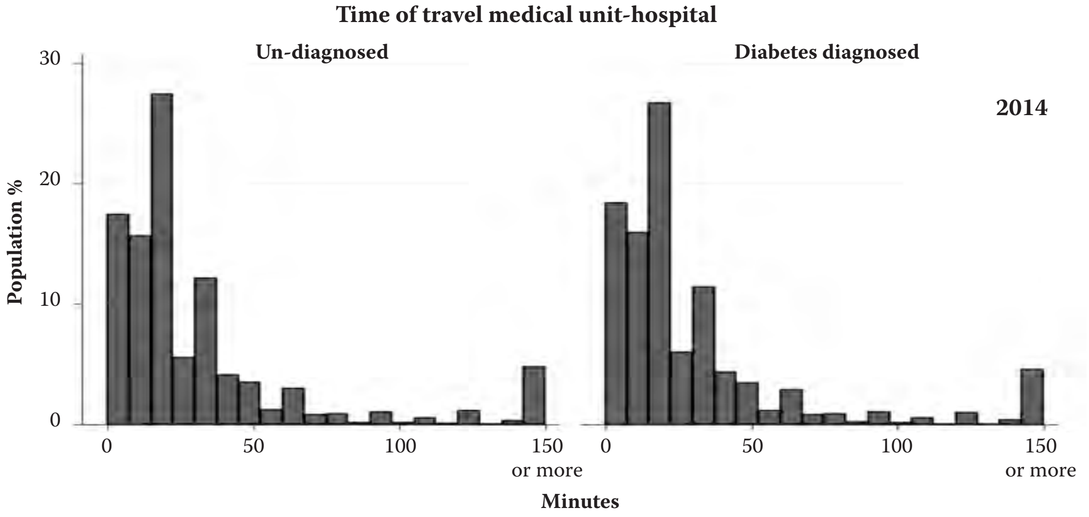
* Any observation over 60 BMI is shown on the last bin.

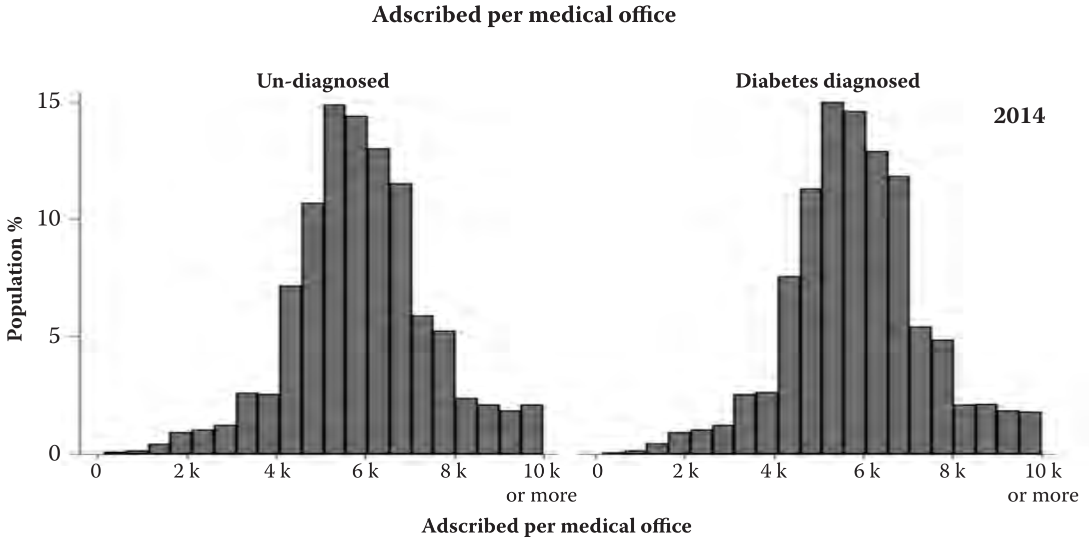




* Any observation of over 1,682.25 Mexican pesos (25 minimum wages) is shown on the last bin.







From these histograms we can see with much more detail the difference that each variable has amongst diagnosed and non-diagnosed patients. We can identify that age and BMI along with capillary glucose appear to be very relevant risk factors and that head of the family's salary, family size, time of travel and population subscribed per medical office present small differences between both populations. Also, since population characteristics do not change much from year to year, we have a better understanding of the disease's correlates. On the next section we describe how we make use of this information in designing the model to be estimated.

The Model

In this section we describe the theory behind the logistic model utilized for estimating the probability of being diagnosed with DM-2. Additionally, we elaborate on how the data is handled to build the model, explain the results obtained and finally display some of the econometric model's predictive power.

Theoretical context

Generalized linear models (GLMs) assume that a dependent variable is explained by some function whose argument can be described linearly. This is a generalization of linear models since the response variable is explained through a link function. An important implication of this type of model is it's effectiveness in predicting specific probabilities. A simple linear model should not be used for this kind of exercise since by construction it is not bound to predict values between 0 and 1. The logit model belongs to the GLM family; in particular it is, along with the probit model, the most common econometric model specification for estimating a model with a binary dependent variable, i.e., a probability.

Logit and probit models belong to the GLM family but have different link functions. A probit model is defined as $pr(Y = 1|X) = \phi(X'\beta)$ where ϕ is the cumulative distribution function of the standard normal distribution. A logit model, on the other hand, is specified as follows: $pr(Y = 1|X) = f(X) = \frac{e^{X'\beta}}{e^{X'\beta+1}}$ where f is known as the logistic function. For both models, marginal effects are not constant, for probit

$$\frac{\partial pr(y = 1|X)}{\partial x_i} = \phi(x)\beta_i$$

and for logit

$$\frac{\partial pr(y = 1|X)}{\partial x_i} = \frac{e^{X'\beta}}{(e^{X'\beta}+1)^2}\beta_i$$

The logistic model has an advantage over the probit model in terms of coefficient interpretation. To see this let's define the odds ratio $OR = \frac{pr(Y=1|X)}{1-pr(Y=1|X)}$, that is how many times it is more likely for Y to be one compared to the probability of Y being 0. In the logistic model $OR = e^{X'\beta}$ which after taking logarithms we obtain $\frac{\partial \log(LR)}{\partial x_i} = \beta_i$. Hence the coefficient β_i can be understood as the effect of a marginal change in x_i over the log-odds ratio. Therefore, the capacity to interpret directly the results led us to the logit functional form.

Given that we are using many factors to estimate a binary response variable, using continuous explanatory variables might prove counterproductive since the relationships between them and diagnosis are hard to pin down. To help with this issue we categorize d all explanatory variables to be able to compare different groups within the population. This technique makes it easier to identify the population at risk highlighting the contrast between a base group and a group that satisfies certain category amongst all variables. This also helped when incorporating interaction terms based on the institute's medical guidelines.

Model Specification

As mentioned above, all explanatory variables were categorized in order to be able to calculate risks for individuals within each group of the population, including interaction terms. For this process we relied heavily on advice from the institute's medical expert staff. Moreover, a similar scheme was used to categorize the corresponding structural variables. Finally, we were able to define the socio-economic, work-related and family categories, as most of them are categorical in nature.

Before going into the interaction terms, we go into the details on how each variable was categorized. It is important to recall that some factors in our model are categorical by nature; hence sex, diagnosis dummies (hypertension, triglycerides and cholesterol), type of worker, insurance type and role in the family enter the model as dummies for each category. On table 3 we explain our categorization for the rest of the variables.

On the basis of the Institute's medical guidelines, risk dummies were defined for the interaction terms on relevant health factors. With respect to Diabetes, these guidelines state that the 4 most relevant risk factors are: hypertension, overweight (BMI>25), age>45 and family history (the existence of a diabetic relative). A capillary glucose risk dummy (measure of over 125 mg/dl, not taken necessarily after fasting) was also defined, since this test is used as a filter to receive the hemoglobin glycation laboratory test at IMSS (taken on an empty stomach) which is used to confirm diagnosis.

Once the main risk factors for developing diabetes were identified, , interactions of every possible combination among them were also entered into the estimation. Therefore, 31 interaction dummies were generated, including, for instance, overweight and over 45 years of age. The intention was to capture the effect of combinations of risk factors play in the prediction of the probability of being diagnosed with diabetes.

Since the effect of certain variables is gender-condition, we separated several dummies by sex. Such was the case for age, BMI, capillary glucose, family history and hypertension as well as their interaction terms, in order to capture how risk factors behave for each population. In summary we will approximate a logit model for predicting the probability of being diagnosed with diabetes by combining 18 variables categorized in a total of 158 dummies. For the actual regression we must exclude one of the dummies for each variable, we exclude the first category in every case. The model can be resumed as follows:

$$\Pr(Y|X) = f(X_{health,\ socio\text{-}economic,\ infrastructure}\ '\beta)$$

Table 3
Variable classification*

Variable	Number of categories	Construction
Age	9	Under 25, 5 year categories between 25 and 60 and over 60.
BMI	7	Under 18.5, between 18.5 and 25, 25 to 27.5, 27.5 to 30, 30 to 35, 35 to 40 and over 40.
Capillary glucose	9	Under 90, 10 unit categories from 90 to 160, from 160 to 200 and over 200.
Diabetic family history	3	No direct relatives with DM-2, one direct relative with DM-2, more than one direct relative with DM-2.
Overweight in family	3	No direct relatives with overweight, one direct relative with overweight in the family, more than one direct relative with overweight.
Obesity in family	3	No obese relatives, one obese relative, more than one obese relative.
Salary	6	We round the salary to the closest number of minimum wages a person receives. We consider 6 or more in the last category.
Family size	6	Number of relatives from 0 to 5, 5 or more compose the last category.
Subscribed population per doctor's office	8	Less than 4,500, categories of 500 from 4,500 to 7,500 and over 7,500.
Time of travel from medical unit to hospital	6	Under 10 minutes, categories of 5 until 30 minutes and over 30 minutes.

* This table reports the discretization of every variable that is not categorical by nature and will be used in our estimations.

where f is the logistic function previously discussed. Once the model design is specified it is possible to adjust the data and contrast the results with what we expected as well as asses it's predictive power to evaluate its impact on IMSS' capacity to detect and diagnose DM-2 patients effectively.

The Results

On this section we will present the results obtained from running the model discussed above on our datasets. As mentioned earlier we ran the logit regression independently for 2012, 2013 and 2014. We start by discussing the coefficients obtained and then move to describe our evaluation of the model's predictive power.

Stability of Estimated Coefficients

On one hand, the coefficients related to the risk factors identified by the institute's clinical guidelines come out positive and statistically significant at the 1% level, with many interactive dummies turning out statistically significant as well. The last categories of time of travel to hospital are positive and also significant, supporting our intuition with respect to this variable in the sense that that the farther away a first level unit (UMF) is from its corresponding hospital of reference the more prone a person subscribed to it is of suffering from diabetes. Finally, family size and the head of the household's salary play a minor but statistically significant role indicating that the bigger the family or higher salaries reduce the chance of a positive diagnosis.

On table 4 the values of the afore-mentioned estimated coefficients for the 3 years, for which the model was run are introduced, as well as the pseudo and the number of observations corresponding to each logit regression. Since the number of categorized variables is large and we are interested in the model's predictive abilities, only the most relevant factors from our point of view are included. The whole set of results including coefficients for each category and for each of the variables that entered into the model can be checked in table 5.

On table 4 all the risk dummies (age, hypertension, diabetic family history, BMI and capillary glucose) turn out positive and statistically significant at the 1% level for both men and women. Dummies for obesity within the family capture some cultural characteristics, with both the presence of a singular obese person and at least two within the family having a positive significant effect, with the latter being larger. Having undergone a preventive checkup reduces substantially the probability of being diagnosed with diabetes. These results highlight that the econometric estimation seems solid from a medical standpoint, as the coefficients are the most relevant determinants of DM-2 risk in accordance with medical guidelines.

Table 4
Summary coefficient results

	2012	2013	2014
Observations	10,550,391	11,575,182	12,310,076
Pseudo R^2	0.25	0.25	0.25
Variable	Coef (SE)	Coef (SE)	Coef (SE)
Man overweight	1.17***	1.13***	1.15***
	(0.06)	(0.06)	(0.06)
Man hypertension	1.01***	1.02***	0.93***
	(0.01)	(0.01)	(0.01)
Man with diabetes diagnosed relative	0.95***	0.91***	0.88***
	(0.01)	(0.01)	(0.01)
Man over 45	2.30***	2.36***	2.38***
	(0.01)	(0.01)	(0.01)
Man with glucose measure over 125 mg/dl	1.50***	1.65***	1.61***
	(0.05)	(0.05)	(0.05)
Man with two or more diabetic relatives	0.22***	0.24***	0.29***
	(0.01)	(0.01)	(0.01)
Woman overweight	1.81***	1.76***	1.70***
	(0.02)	(0.02)	(0.02)

Woman hypertension	1.12***	1.11***	1.04***
	(0.01)	(0.01)	(0.01)
Woman with diabetes diagnosed relative	0.84***	0.81***	0.75***
	(0.01)	(0.01)	(0.01)
Woman over 45	2.70***	2.73***	2.69***
	(0.01)	(0.01)	(0.01)
Woman with glucose measure over 125 mg/dl	1.47***	1.51***	1.61***
	(0.04)	(0.04)	(0.05)
Woman with two or more diabetic relatives	0.20***	0.23***	0.28***
	(0.01)	(0.01)	(0.01)
Permanent worker	0.23***	0.23***	0.23***
	(0.01)	(0.01)	(0.01)
Relative to permanent worker	0.06***	0.06***	0.08***
	(0)	(0)	(0)
One obese person in the family	0.10***	0.11***	0.11***
	(0)	(0)	(0)
At least 2 obese persons in the family	0.16***	0.16***	0.16***
	(0.01)	(0.01)	(0.01)
Student's insurance	-1.86***	-1.89***	-1.93***
	(0.04)	(0.04)	(0.03)
Between 25 and 30 minutes from medical unit to hospital	0.11***	0.08***	0.08***
	(0)	(0)	(0)
More than 30 minutes from medical unit to hospital	0.01***	0.01*	0.02***
	(0)	(0)	(0)
Preventive checkup	-1.71***	-1.47***	-1.60***
	(0.01)	(0.01)	(0.01)
1 year experience	-0.72***	-0.61***	-0.53***
	(0.03)	(0.02)	(0.02)
Between 1 and 5 years experience	-0.70***	-0.53***	-0.39***
	(0.01)	(0.01)	(0.01)
Married partner dummy	0.79***	0.84***	0.96***
	(0.02)	(0.02)	(0.02)
Family head dummy	0.14***	0.05	0.31***
	(0.03)	(0.03)	(0.02)
Father of family head dummy	0.86***	0.91***	1.04***
	(0.02)	(0.02)	(0.02)
Mother of family head dummy	1.07***	1.11***	1.25***
	(0.02)	(0.02)	(0.02)
At last 6 minimum wages earned	-0.19***	-0.19***	-0.17***
	(0)	(0)	(0)

(*), (**), (***) denote significance at 10, 5 and 1 percent levels respectively.

From the socio-economic, work-related and family variables we can see that being a permanent worker increases the probability of being diagnosed with diabetes, and belonging to a permanent worker's family does so as well but with a much smaller effect. Also, we can see that workers that have little experience on the formal sector are less prone to receiving a DM-2 diagnosis. Being the main breadwinner within the family increases significantly the probability of being diagnosed relative to being an offspring. Moreover, it is interesting to see that being the married partner of the worker has a larger effect and, as expected, being a parent has the largest effect, with the mother's effect being slightly bigger. Having a larger salary reduces the probability of being diagnosed with DM-2. Finally, being ensured as a student reduces the diagnosis probability significantly.

We can see that being subscribed to a medical unit that is between a 25 and 30 minute drive to its hospital of reference increases substantially the probability of diagnosis, relative to attending a medical unit within the hospital complex, while being farther away than a 30 minute drive upholds risk in a smaller proportion. This could be explained by the fact that far and away medical units tend to mostly rural population whose active lifestyles make them less prone to obesity and diabetes, reducing the effect of lack of access to medical care on the risk of diagnosis.

Table 5
Full coefficient results

	2012	**2013**	**2014**
Observations	10,550,391	11,575,182	12,310,076
Pseudo R^2	0.25	0.25	0.25
Variable	Coef (SE)	Coef (SE)	Coef (SE)
Man overweight	1.17***	1.13***	1.15***
	(0.06)	(0.06)	(0.06)
Man hypertension	1.01***	1.02***	0.93***
	(0.01)	(0.01)	(0.01)
Man with diabetes diagnosed relative	0.95***	0.91***	0.88***
	(0.01)	(0.01)	(0.01)
Man over 45 years old	2.30***	2.36***	2.38***
	(0.01)	(0.01)	(0.01)
Man overweight and hypertension	-0.21***	-0.22***	-0.22***
	(0.01)	(0.01)	(0.01)
Man overweight and with a diabetes diagnosed relative	-0.02	0	0
	(0.01)	(0.01)	(0.01)
Man overweight over 45 years old	-0.29***	-0.28***	-0.27***
	(0.01)	(0.01)	(0.01)
Man over 45 years with hypertension	-0.90***	-0.95***	-0.95***
	(0.01)	(0.01)	(0.01)

Man over 45 years old with a diabetic relative	-0.35***	-0.35***	-0.36***
	(0.01)	(0.01)	(0.01)
Man with hypertension and a diabetic relative	-0.20***	-0.20***	-0.16***
	(0.01)	(0.01)	(0.01)
Man with high cholesterol	0.43***	0.39***	0.30***
	(0.02)	(0.02)	(0.02)
Man with high tryglicerides	-0.55***	-0.59***	-0.58***
	(0.02)	(0.02)	(0.02)
Woman overweight	1.81***	1.76***	1.70***
	(0.02)	(0.02)	(0.02)
Woman hypertension	1.12***	1.11***	1.04***
	(0.01)	(0.01)	(0.01)
Woman with diabetes diagnosed relative	0.84***	0.81***	0.75***
	(0.01)	(0.01)	(0.01)
Woman over 45 years old	2.70***	2.73***	2.69***
	(0.01)	(0.01)	(0.01)
Woman overweight and hypertension	-0.20***	-0.22***	-0.22***
	(0.01)	(0.01)	(0.01)
Woman overweight and with a diabetes diagnosed relative	-0.14***	-0.14***	-0.12***
	(0.01)	(0.01)	(0.01)
Woman overweight over 45 years old	-0.55***	-0.52***	-0.49***
	(0.01)	(0.01)	(0.01)
Woman over 45 years with hypertension	-1.21***	-1.23***	-1.24***
	(0.01)	(0.01)	(0.01)
Woman over 45 years old with a diabetic relative	-0.10***	-0.12***	-0.10***
	(0.01)	(0.01)	(0.01)
Woman with hypertension and a diabetic relative	-0.19***	-0.19***	-0.16***
	(0.01)	(0.01)	(0.01)
Woman with high cholesterol	0.71***	0.65***	0.55***
	(0.02)	(0.02)	(0.01)
Woman with high tryglicerides	-0.48***	-0.51***	-0.54***
	(0.01)	(0.01)	(0.01)
Permanent worker	0.23***	0.23***	0.23***
	(0.01)	(0.01)	(0.01)
Relative to permanent worker	0.06***	0.06***	0.08***
	(0)	(0)	(0)
Woman with 2 or more diabetic relatives	0.20***	0.23***	0.28***
	(0.01)	(0.01)	(0.01)
Man with 2 or more diabetic relatives	0.22***	0.24***	0.29***
	(0.01)	(0.01)	(0.01)

City worker	-0.02	0	-0.18***
	(0.02)	(0.02)	(0.01)
Rural worker	1.00***	1.02***	-0.20***
	(0.03)	(0.03)	(0.02)
Relative of an IMSS or CFE (Compañía Federal de Electricidad) worker or a student	0.66***	0.71***	0.81***
	(0.03)	(0.03)	(0.03)
Relative to informal worker with family insurance	0.17***	0.19***	0.49***
	(0.02)	(0.02)	(0.03)
Government workers	0.13***	0.16***	-0.03
	(0.02)	(0.02)	(0.02)
Other kind of worker	0.30***	0.32***	0.14***
	(0.02)	(0.02)	(0.02)
Not reported kind of worker	0.48***	0.48***	0.27***
	(0.02)	(0.02)	(0.02)
Man with 18.5<=BMI<25	0.51***	0.48***	0.49***
	(0.02)	(0.02)	(0.02)
Man with 25<=BMI<27.5	-0.11*	-0.12*	-0.13*
	(0.06)	(0.05)	(0.05)
Man with 27.5<=BMI<30	-0.01	-0.02	-0.03
	(0.06)	(0.05)	(0.05)
Man with 30<=BMI<35	0.11*	0.11*	0.09
	(0.06)	(0.05)	(0.05)
Man with 35<=BMI<40	0.29***	0.28***	0.29***
	(0.06)	(0.05)	(0.05)
Man with 40<=BMI	0.43***	0.42***	0.44***
	(0.06)	(0.05)	(0.05)
Woman with 18.5<=BMI<25	0.58***	0.56***	0.49***
	(0.02)	(0.02)	(0.01)
Woman with 25<=BMI<27.5	-0.46***	-0.45***	-0.47***
	(0.01)	(0.01)	(0.01)
Woman with 27.5<=BMI<30	-0.35***	-0.34***	-0.36***
	(0.01)	(0.01)	(0.01)
Woman with 30<=BMI<35	-0.21***	-0.19***	-0.20***
	(0.01)	(0.01)	(0.01)
Woman with 35<=BMI<40	-0.04***	-0.02	-0.03**
	(0.01)	(0.01)	(0.01)
Woman with 40<=BMI	0.16***	0.18***	0.18***
	(0.01)	(0.01)	(0.01)
Man with 25<age<=30	-0.82***	-0.84***	-0.86***
	(0.01)	(0.01)	(0.01)

Man with 25<age<=30	-0.17***	-0.20***	-0.20***
	(0.01)	(0.01)	(0.01)
Man with 30<age<=35	0.34***	0.34***	0.35***
	(0.01)	(0.01)	(0.01)
Man with 35<age<=40	0.64***	0.65***	0.68***
	(0.01)	(0.01)	(0.01)
Man with 40<age<=45	-0.42***	-0.42***	-0.41***
	(0.01)	(0.01)	(0.01)
Man with 45<age<=50	0.05***	0.05***	0.08***
	(0.01)	(0.01)	(0.01)
Man with 50<age<=55	0.17***	0.18***	0.22***
	(0.01)	(0.01)	(0.01)
Man with 55<age<=60	0.13***	0.13***	0.18***
	(0.01)	(0.01)	(0.01)
Woman with 25<age<=30	-0.72***	-0.71***	-0.70***
	(0.01)	(0.01)	(0.01)
Woman with 25<age<=30	-0.15***	-0.15***	-0.18***
	(0.01)	(0.01)	(0.01)
Woman with 30<age<=35	0.29***	0.29***	0.27***
	(0.01)	(0.01)	(0.01)
Woman with 35<age<=40	0.56***	0.56***	0.55***
	(0.01)	(0.01)	(0.01)
Woman with 40<age<=45	-0.52***	-0.55***	-0.56***
	(0.01)	(0.01)	(0.01)
Woman with 45<age<=50	0	-0.01*	0
	(0.01)	(0.01)	(0.01)
Woman with 50<age<=55	0.16***	0.15***	0.17***
	(0.01)	(0)	(0)
Woman with 55<age<=60	0.33***	0.34***	0.36***
	(0.01)	(0.01)	(0.01)
Woman	-0.59***	-0.60***	-0.46***
	(0.03)	(0.03)	(0.02)
Worker	0.14***	0.05	0.31***
	(0.03)	(0.03)	(0.02)
Spouse	0.79***	0.84***	0.96***
	(0.02)	(0.02)	(0.02)
Partner (living together)	0.74***	0.78***	0.89***
	(0.02)	(0.02)	(0.02)
Father	0.86***	0.91***	1.04***
	(0.02)	(0.02)	(0.02)

Mother	1.07***	1.11***	1.25***
	(0.02)	(0.02)	(0.02)
Family of 2	-0.02***	-0.01***	-0.01***
	(0)	(0)	(0)
Family of 3	-0.12***	-0.10***	-0.09***
	(0)	(0)	(0)
Family of 4	-0.16***	-0.14***	-0.13***
	(0)	(0)	(0)
Family of 5	-0.16***	-0.13***	-0.13***
	(0.01)	(0)	(0)
Family of 6	-0.15***	-0.13***	-0.13***
	(0.01)	(0.01)	(0.01)
1 other overweight person in the family	-0.05***	-0.03***	-0.04***
	(0)	(0)	(0)
2 or more overweight person in the family	-0.16***	-0.14***	-0.15***
	(0)	(0)	(0)
1 other obese person in the family	0.10***	0.11***	0.11***
	(0)	(0)	(0)
2 or more other obese persons in the family	0.16***	0.16***	0.16***
	(0.01)	(0.01)	(0.01)
One year of formal work experience	-0.72***	-0.61***	-0.53***
	(0.03)	(0.02)	(0.02)
Between 2 and 5 years of formal work experience	-0.70***	-0.53***	-0.39***
	(0.01)	(0.01)	(0.01)
Between 6 and 10 years of formal work experience	-0.37***	-0.14***	0.02
	(0.01)	(0.01)	(0.01)
Between 10 and 15 years of formal work experience	-0.04***	0.11***	0.19***
	(0.01)	(0.01)	(0.01)
Between 15 and 20 years of formal work experience	-0.02	0.12***	0.21***
	(0.01)	(0.01)	(0.01)
Over 20 years of formal work experience	0	0.15***	0.21***
	(0.01)	(0.01)	(0.01)
Student's insurance	-1.86***	-1.89***	-1.93***
	(0.04)	(0.04)	(0.03)
Pension insurance	0.13***	0.16***	0.22***
	(0.01)	(0)	(0)
Purchased insurance	0.05	0	0.02
	(0.03)	(0.02)	(0.02)
Other insurance	-0.13***	-0.18***	-0.25***
	(0.03)	(0.02)	(0.02)

Salary close to 2 minimum wages	-0.05***	-0.06***	-0.06***
	(0)	(0)	(0)
Salary close to 3 minimum wages	-0.09***	-0.09***	-0.09***
	(0)	(0)	(0)
Salary close to 4 minimum wages	-0.12***	-0.11***	-0.10***
	(0)	(0)	(0)
Salary close to 5 minimum wages	-0.12***	-0.11***	-0.10***
	(0.01)	(0.01)	(0.01)
Salary over 6 minimum wages	-0.19***	-0.19***	-0.17***
	(0)	(0)	(0)
Adscribed per medical office between 4,500 and 5,000	-0.03***	-0.02***	0.01***
	(0)	(0)	(0)
Adscribed per medical office between 5,000 and 5,500	-0.01*	0.01***	0.01*
	(0)	(0)	(0)
Adscribed per medical office between 5,500 and 6,000	-0.01**	0.02***	0.02***
	(0)	(0)	(0)
Adscribed per medical office between 6,000 and 6,500	-0.02***	0	0.01**
	(0)	(0)	(0)
Adscribed per medical office between 6,500 and 7,000	-0.01*	0.01	0.02***
	(0)	(0)	(0)
Adscribed per medical office between 7,000 and 7,500	-0.01*	0.02***	0.04***
	(0)	(0)	(0)
Adscribed per medical office between 7,500 and 8,000	-0.01**	0.01**	0.02***
	(0)	(0)	(0)
Between 10 and 15 minutes transport from first level medical unit to hospital	0.03***	0.01***	0.01***
	(0)	(0)	(0)
Between 15 and 20 minutes transport from first level medical unit to hospital	0.03***	0.02***	0.04***
	(0)	(0)	(0)
Between 20 and 25 minutes transport from first level medical unit to hospital	0.02***	0.01**	0.03***
	(0)	(0)	(0)
Between 25 and 30 minutes transport from first level medical unit to hospital	0.11***	0.08***	0.08***
	(0)	(0)	(0)
Over 30 minutes transport from first level medical unit to hospital	0.01***	0.01*	0.02***
	(0)	(0)	(0)
First level medical unit in Baja California	0.31***	0.39***	0.42***
	(0.01)	(0.01)	(0.01)
First level medical unit in Baja California Sur	0.12***	0.24***	0.22***
	(0.01)	(0.01)	(0.01)
First level medical unit in Campeche	0.04**	0.17***	0.10***
	(0.01)	(0.01)	(0.01)

First level medical unit in Coahuila	0.21***	0.29***	0.27***
	(0.01)	(0.01)	(0.01)
First level medical unit in Colima	0.11***	0.12***	0.13***
	(0.01)	(0.01)	(0.01)
First level medical unit in Chiapas	0.02	0.10***	0.12***
	(0.01)	(0.01)	(0.01)
First level medical unit in Chihuahua	0.06***	0.14***	0.17***
	(0.01)	(0.01)	(0.01)
First level medical unit in Distrito Federal	0.10***	0.18***	0.18***
	(0.01)	(0.01)	(0.01)
First level medical unit in Durango	0.06***	0.22***	0.15***
	(0.01)	(0.01)	(0.01)
First level medical unit in Guanajuato	0.19***	0.21***	0.20***
	(0.01)	(0.01)	(0.01)
First level medical unit in Guerrero	0.13***	0.23***	0.19***
	(0.01)	(0.01)	(0.01)
First level medical unit in Hidalgo	0.14***	0.20***	0.23***
	(0.01)	(0.01)	(0.01)
First level medical unit in Jalisco	0.05***	0.12***	0.16***
	(0.01)	(0.01)	(0.01)
First level medical unit in Estado de México	0.20***	0.27***	0.23***
	(0.01)	(0.01)	(0.01)
First level medical unit in Michoacán	0.07***	0.12***	0.14***
	(0.01)	(0.01)	(0.01)
First level medical unit in Morelos	0.15***	0.22***	0.24***
	(0.01)	(0.01)	(0.01)
First level medical unit in Nayarit	0.03*	0.09***	0.20***
	(0.01)	(0.01)	(0.01)
First level medical unit in Nuevo León	0.37***	0.49***	0.49***
	(0.01)	(0.01)	(0.01)
First level medical unit in Oaxaca	0.05***	0.15***	0.20***
	(0.01)	(0.01)	(0.01)
First level medical unit in Puebla	0.14***	0.21***	0.17***
	(0.01)	(0.01)	(0.01)
First level medical unit in Querétaro	0.09***	0.14***	0.11***
	(0.01)	(0.01)	(0.01)
First level medical unit in Quintana Roo	-0.08***	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
First level medical unit in San Luis Potosí	0.23***	0.39***	0.48***
	(0.01)	(0.01)	(0.01)

First level medical unit in Sinaloa	0.05***	0.11***	0.02
	(0.01)	(0.01)	(0.01)
First level medical unit in Sonora	0.09***	0.18***	0.20***
	(0.01)	(0.01)	(0.01)
First level medical unit in Tabasco	0.08***	0.16***	0.21***
	(0.01)	(0.01)	(0.01)
First level medical unit in Tamaulipas	0.22***	0.35***	0.37***
	(0.01)	(0.01)	(0.01)
First level medical unit in Tlaxcala	0.07***	0.12***	0.13***
	(0.01)	(0.01)	(0.01)
First level medical unit in Veracruz	0.10***	0.19***	0.20***
	(0.01)	(0.01)	(0.01)
First level medical unit in Yucatán	0.10***	0.19***	0.23***
	(0.01)	(0.01)	(0.01)
First level medical unit in Zacatecas	-0.05***	0.07***	0.06***
	(0.01)	(0.01)	(0.01)
Registered for morning service	0.25***	0.26***	0.38***
	(0.01)	(0.01)	(0.01)
Registered for afternoon service	0.20***	0.21***	0.34***
	(0.01)	(0.01)	(0.01)
Man with capillary glucose measurement of under 90 mg/dl	0.02	0.05***	0.02
	(0.01)	(0.01)	(0.01)
Man with capillary glucose measurement between 90 mg/dl and 110 mg/dl	0.09***	-0.20***	-0.15***
	(0.01)	(0.01)	(0.01)
Man with capillary glucose measurement between 110 mg/dl and 120 mg/dl	0.51***	0.18***	0.16***
	(0.01)	(0.01)	(0.01)
Man with capillary glucose measurement between 120 mg/dl and 130 mg/dl	1.11***	0.76***	0.72***
	(0.01)	(0.01)	(0.01)
Man with capillary glucose measurement between 120 mg/dl and 130 mg/dl	1.51***	1.12***	1.07***
	(0.03)	(0.03)	(0.03)
Man with capillary glucose measurement between 130 mg/dl and 140 mg/dl	2.04***	1.78***	1.81***
	(0.03)	(0.03)	(0.03)
Man with capillary glucose measurement between 140 mg/dl and 150 mg/dl	2.33***	2.00***	1.98***
	(0.03)	(0.03)	(0.03)
Man with capillary glucose measurement between 150 mg/dl and 160 mg/dl	2.62***	2.30***	2.34***
	(0.03)	(0.03)	(0.03)
Man with capillary glucose measurement over 160 mg/dl	3.63***	3.32***	3.36***
	(0.03)	(0.03)	(0.03)
Woman with capillary glucose measurement between 90 mg/dl and 110 mg/dl	-0.03***	-0.28***	-0.19***
	(0.01)	(0.01)	(0.01)

Woman with capillary glucose measurement between 110 mg/dl and 120 mg/dl	0.41***	0.11***	0.11***
	(0.01)	(0.01)	(0.01)
Woman with capillary glucose measurement between 120 mg/dl and 130 mg/dl	1.06***	0.71***	0.71***
	(0.01)	(0.01)	(0.01)
Woman with capillary glucose measurement between 120 mg/dl and 130 mg/dl	1.49***	1.12***	1.07***
	(0.02)	(0.02)	(0.02)
Woman with capillary glucose measurement between 130 mg/dl and 140 mg/dl	2.15***	1.82***	1.91***
	(0.02)	(0.02)	(0.02)
Woman with capillary glucose measurement between 140 mg/dl and 150 mg/dl	2.52***	2.10***	2.15***
	(0.02)	(0.02)	(0.03)
Woman with capillary glucose measurement between 150 mg/dl and 160 mg/dl	2.85***	2.46***	2.51***
	(0.02)	(0.02)	(0.02)
Woman with capillary glucose measurement over 160 mg/dl	3.65***	3.28***	3.29***
	(0.02)	(0.02)	(0.02)
Preventive check-up (PrevenIMSS)	-1.71***	-1.47***	-1.60***
	(0.01)	(0.01)	(0.01)
Man with capillary glucose measurement of over 125 mg/dl	1.50***	1.65***	1.61***
	(0.05)	(0.05)	(0.05)
Man overweight and with capillary glucose measurement of over 125 mg/dl	-0.55***	-0.63***	-0.56***
	(0.04)	(0.04)	(0.05)
Man hypertension and with capillary glucose measurement of over 125 mg/dl	-0.69***	-0.75***	-0.69***
	(0.05)	(0.05)	(0.06)
Man with diabetes diagnosed relative and with capillary glucose measurement of over 125 mg/dl	-0.52***	-0.47***	-0.44***
	(0.06)	(0.06)	(0.06)
Man over 45 years old and with capillary glucose measurement of over 125 mg/dl	-1.49***	-1.66***	-1.73***
	(0.04)	(0.04)	(0.05)
Man overweight and hypertension and with capillary glucose measurement of over 125 mg/dl	0.06	0.11**	0.10**
	(0.04)	(0.04)	(0.04)
Man overweight and with a diabetes diagnosed relative and with capillary glucose measurement of over 125 mg/dl	0.08	0.01	-0.07
	(0.04)	(0.04)	(0.04)
Man overweight over 45 years old and with capillary glucose measurement of over 125 mg/dl	0.38***	0.47***	0.46***
	(0.05)	(0.05)	(0.05)
Man over 45 years with hypertension and with capillary glucose measurement of over 125 mg/dl	0.56***	0.61***	0.56***
	(0.05)	(0.05)	(0.05)
Man over 45 years old with a diabetic relative and with capillary glucose measurement of over 125 mg/dl	0.30***	0.32***	0.39***
	(0.05)	(0.05)	(0.06)
Man with hypertension and a diabetic relative and with capillary glucose measurement of over 125 mg/dl	0.05	0.08*	0.08*
	(0.03)	(0.03)	(0.04)

Woman with capillary glucose measurement of over 125 mg/dl	1.47***	1.51***	1.61***
	(0.04)	(0.04)	(0.05)
Woman overweight and with capillary glucose measurement of over 125 mg/dl	-0.52***	-0.47***	-0.55***
	(0.04)	(0.04)	(0.04)
Woman hypertension and with capillary glucose measurement of over 125 mg/dl	-0.69***	-0.64***	-0.68***
	(0.05)	(0.05)	(0.05)
Woman with diabetes diagnosed relative and with capillary glucose measurement of over 125 mg/dl	-0.30***	-0.31***	-0.22***
	(0.06)	(0.06)	(0.06)
Woman over 45 years old and with capillary glucose measurement of over 125 mg/dl	-1.41***	-1.40***	-1.55***
	(0.04)	(0.04)	(0.05)
Woman overweight and hypertension and with capillary glucose measurement of over 125 mg/dl	0.07*	0.06	0.08*
	(0.03)	(0.03)	(0.03)
Woman overweight and with a diabetes diagnosed relative and with capillary glucose measurement of over 125 mg/dl	-0.02	0.03	-0.09*
	(0.04)	(0.04)	(0.05)
Woman overweight over 45 years old and with capillary glucose measurement of over 125 mg/dl	0.36***	0.27***	0.36***
	(0.04)	(0.04)	(0.05)
Woman over 45 years with hypertension and with capillary glucose measurement of over 125 mg/dl	0.49***	0.48***	0.51***
	(0.04)	(0.04)	(0.04)
Woman over 45 years old with a diabetic relative and with capillary glucose measurement of over 125 mg/dl	0.18***	0.18***	0.18***
	(0.04)	(0.05)	(0.05)
Woman with hypertension and a diabetic relative and with capillary glucose measurement of over 125 mg/dl	0.12***	0.05	0.09**
	(0.03)	(0.03)	(0.03)
Constant	-3.89***	-4.03***	-4.38***
	(0.04)	(0.04)	(0.04)

(*), (**), (***) denote significance at 10, 5 and 1 percent levels respectively.

Once we have reviewed the coefficients resulting from the estimation and are certain of their consistency, we turn to study the most relevant result, the predicted probability. When analyzing this for the two groups –diagnosed and un-diagnosed beneficiaries– the model's predictive powers can be detected. On table 6 the mean predicted probability is displayed, together with standard deviation for diagnosed and un-diagnosed patients for each year.

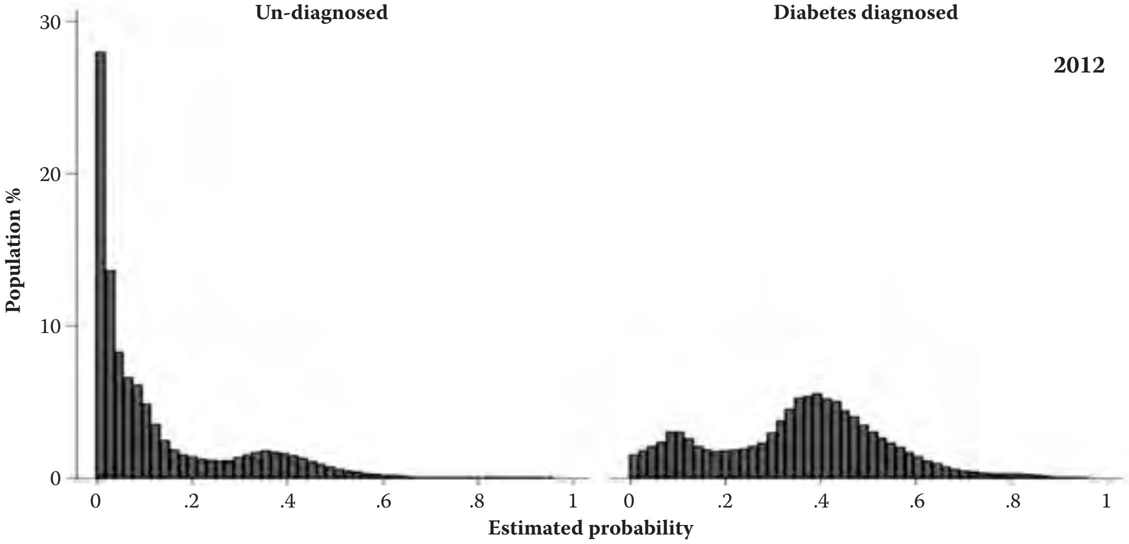
Table 6
Mean predicted probability*

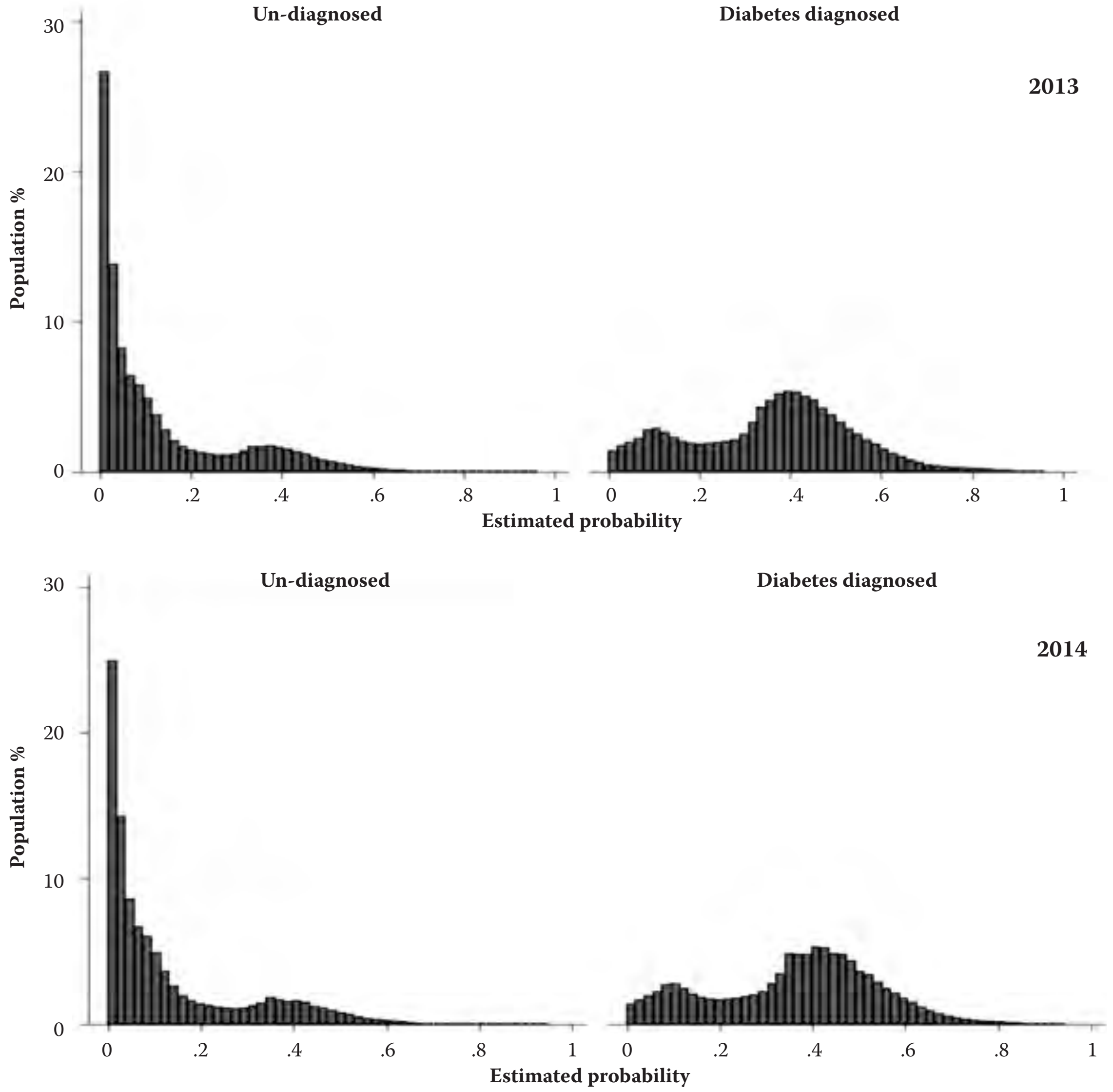
	2012		**2013**		**2014**	
Mean estimated probability (*standard deviation*)	0.16 (0.18)		0.17 (0.18)		0.17 (0.18)	
Population	**Not diagnosed**	**Diagnosed diabetics**	**Not diagnosed**	**Diagnosed diabetics**	**Not diagnosed**	**Diagnosed diabetics**
Mean estimated probability (*standard deviation*)	0.13 (0.15)	0.36 (0.18)	0.13 (0.15)	0.36 (0.18)	0.13 (0.16)	0.36 (0.18)

* The first row of this table uses all the population estimated on each year. For the second row we separate diagnosed patients from others and estimate their respective mean probabilities, for each year.

As evidence of the model's predictive power we observe that the probability is much higher for diagnosed patients. Moreover, for each of the three years, over 81% of patients diagnosed with diabetes have a predicted probability above the mean and 58% above the mean plus one standard deviation. To provide more insight into these results, we display on figure 10 the histograms of the predicted probability for each group of the population for each year. On this figure the vertical axis captures the share of the population and the horizontal axis contains the predicted probability.

Figure 10
Estimated probability of diagnosis





These charts illustrate the model's findings quite clearly. First of all, the diagnosed population features a semi-normal distribution forming around the mean. This is interpreted as evidence that not every diabetic patient must present all risk factors, with some individuals being diabetic due to unknown characteristics. It can also be recognized that the model identifies a large amount of un-diagnosed patients as healthy, with less than 0.05 probability. Moreover, to the right of the mass containing these healthy individuals, a replica of the shape of the histogram resulting for the diagnosed population emerges. These features are present in each year considered (which are all estimated independently), and thus highlight the model's statistical stability over time.

Evaluating the Model's Predictive Power

The results obtained so far point towards model stability, which clearly is derived in a significant extent from the large size of the population we analyze. In other words, we feel pretty confident about the certainty of the estimated coefficients up to this point and recognize that this outcome would not have been feasible had the sample been considerably smaller in size.

The next natural step is to evaluate the possibility of leveraging the estimated risk-profiling methodology and utilize it to predict disease outcomes appearing in our datasets. That is, we would ask the question: can a given level of risk –as indicated by the model– lead us to predict the probability of being diagnosed with diabetes in a previous or subsequent year? Were this to be the case we could confirm that our estimated model portrays predictive power over time, allowing us to utilize the latest estimation (2014) to make inferences on risk for out-of-sample IMSS beneficiaries.

Table 7 displays results for differences between predictions for 2012 and 2013 using each year's corresponding forecast and the one obtained from the 2014 model on the other years datasets. The outcomes not only bare great resemblance, but vary more the further away from the year the predictive model was adjusted. This similarity can justify the utilization of the most recent estimation (2014) to predict the chance of a positive diagnosis for those individuals not within our database. As a corollary, it begs to notice that the statistical inference we just obtained –our main econometric estimation– needs to be re-calibrated every so often in order to update its validity and to be able to capture more recent fluctuations in the spread of the disease. We would go so far as to venture the idea that our econometric tool would allow health authorities to have a better handle on the spread of a disease that seemingly is behaving somewhat like and epidemic.

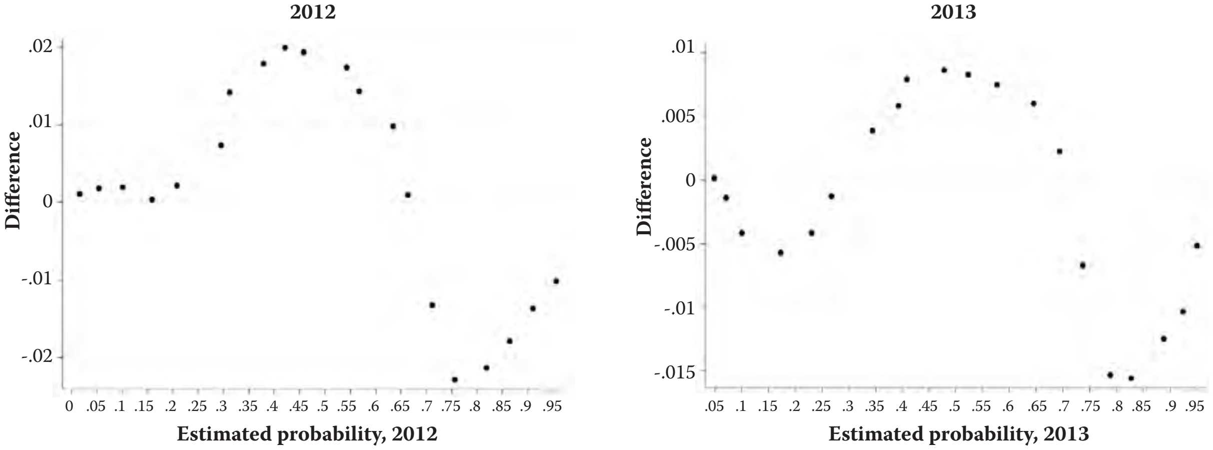
Table 7
Mean difference on prediction*

Year	Observations	2014 prediction difference mean (SD)	Observations with a difference larger than 0.1 (% of total)	Observations with a difference larger than 0.05 (% of total)	Observations with a difference larger than 0.01 (% of total)
2012	10,550,391	0.005 (0.018)	1,559 (0%)	360,514 (3%)	3,605,628 (34%)
2013	11,575,182	0.0005 (0.011)	15 (0%)	26,760 (0.2%)	2,761,594 (24%)

* On this table we report the mean difference in prediction from using the 2014 data adjusted model to predict risk on 2012 and 2013 relative to adjusting a model for each of these years separately.

To strengthen confidence on using the 2014 adjustment as a base model, it is worthwhile making sure that the differences in estimation were not concentrated on a specific level of predicted probability. Figure 11 displays a scatterplot of the mean difference in predicted probability for each year, with the vertical axis representing the difference in probability and the horizontal axis containing the original predicted probability for each year. We plotted the mean difference in 0.05 intervals in order to be able to highlight the estimation results.

Figure 11
Mean difference in estimation



The fact that the outcomes do not vary much along the whole range of prediction provides reassurance about the ability to utilize the 2014-model as a base to assign a risk profile to out of sample observations, including recent entries. These results underline once more the statistical stability of our original estimation. We therefore can feel confident with the results so far and proceed to generate a nominative list of 17.5 million IMSS beneficiaries that assigns to each one of them a predicted risk profile of positive DM-2 diagnosis. In order to do so, whenever an individual has more than one, the most recent prediction is kept. The ensuing list is described in table 8.

Table 8 depicts our risk-profiling scheme for potential Diabetics at IMSS. As can be observed, the proportion of diagnosed diabetics in each segment of the population rises with the estimated probability, implying that the model does have the ability to classify risk-prone patients with high certainty. Also, we can see that the percentage of diagnosed diabetics in each decile is close to the center of the probability segment, which highlights once more the model's good statistical fit. In our view, the information contained in this table represents a very powerful tool for preventive health policy, as it can be utilized as a detection aid to capture more effectively those DM-2 prone individuals within the population of IMSS' beneficiaries.

Table 8
Population distribution by deciles*

Variable	Not diagnosed	Diagnosed with DM2
Mean predicted probability (SD)	0.35 *(0.18)*	0.11 *(0.15)*
Pr<0.1	9,862,678 (97%)	314,311 (3%)
0.1≤Pr<0.2	1,870,153 (86%)	300,841 (14%)
0.2≤Pr<0.3	819,720 (76%)	255,606 (24%)
0.3≤Pr<0.4	1,032,064 (65%)	546,355 (35%)
0.4≤Pr<0.5	840,391 (57%)	840,391 (43%)
0.5≤Pr<0.6	330,429 (48%)	360,673 (52%)
0.6≤Pr<0.7	87,943 (39%)	138,888 (61%)
0.7≤Pr<0.8	16,928 (31%)	37,165 (69%)
0.8≤Pr	3,756 (24%)	11,610 (76%)
Total	14,864,062 (85%)	2,608,032 (15%)

* On this table we show the number and percentage, in parenthesis, of diagnosed patients inside each decile of estimated probability in the model.

Dynamic Validation of Risk-Profiling Process

In order to once again test the model's predictive power our databases lend themselves to let us follow those un-diagnosed individuals on a given year that were diagnosed in a subsequent year. Therefore, we account for the group of IMSS beneficiaries that attended a medical unit (UMF) on a given (base) year not having been diagnosed with DM-2. The main idea is to test the proportion of such patients that were positively diagnosed on a subsequent year.

The exercise can be divided in two cases: one where we test for diagnosis within two years; and one where we test for diagnosis occurring within a year. Thus, departing from 2012 and into 2013 or 2014, which would be the exercise accounting for two years; or, the one year case where we depart from 2012 with diagnosis in 2013, or departing from 2013 with diagnosis in 2014. The results from this exercise are presented in table 9, where we display decile comparisons. To construct this table we use the risk estimation of the base year.

Table 9
New diagnosis by deciles*

Classification	2012-2013		2013-2014		2012-2014	
	Individuals without diagnosis on 2012 (percent of not diagnosed)	Diagnosed on 2013 (percent diagnosed from risk segment)	Individuals without diagnosis on 2013 (percent of not diagnosed)	Diagnosed on 2014 (percent diagnosed from risk segment)	Individuals without diagnosis on 2012 (percent of not diagnosed)	Diagnosed on 2013 or 2014 (percent diagnosed from risk segment)
Pr<0.1	5,557,450 (63%)	46,674 (0.8%)	6,039,699 (63%)	50,527 (0.8%)	5,557,450 (63%)	88,339 (1.6%)
0.1≤Pr<0.2	1,250,078 (14%)	38,327 (3%)	1,376,812 (14%)	40,739 (3%)	1,250,078 (14%)	68,291 (5.4%)
0.2≤Pr<0.3	495,459 (6%)	19,229 (3.8%)	551,586 (6%)	20,828 (3.8%)	495,459 (6%)	68,291 (6.5%)
0.3≤Pr<0.4	654,984 (6%)	29,786 (4.5%)	728,420 (8%)	32,521 (4.5%)	654,984 (6%)	48,451 (7.4%)
0.4≤Pr<0.5	568,779 (6%)	31,348 (5.5%)	613,391 (6%)	32,982 (5.4%)	568,779 (6%)	50,171 (8.8%)
0.5≤Pr<0.6	227,634 (3%)	15,209 (6.7%)	241,290 (3%)	15,717 (6.5%)	227,634 (3%)	24,027 (10.5%)
0.6≤Pr<0.7	63,342 (1%)	5,430 (8.6%)	65,439 (1%)	5,361 (8.2%)	63,342 (1%)	8,199 (12.9%)
0.7≤Pr<0.8	13,508 (0.2%)	2, 117 (15.7%)	13,451 (0.1%)	1,926 (14.3%)	13,508 (0.2%)	3,019 (22.3%)
0.8≤Pr	3,093 (0.04%)	736 (23.8%)	2,973 (0.03%)	628 (21.1%)	3,093 (0.04%)	1,019 (32.9%)
Total population	8, 834,327 (100%)	188,856 (2.1%)	9,633,061 (100%)	201,229 (2.1%)	8,834,327 (100%)	323,991 (3.6%)

* On this table we present the number of individuals, percentage in parenthesis, without diagnosis on each decile of the estimated probability for that base year and the number of this individuals that were diagnosed in the following years, percentage in parenthesis.

As can be recognized from these results, the estimated probability model does have predictive power from 2012 to 2014, as more than half the population comes out with a chance of being diagnosed with diabetes below 2% after 2 years. Also, within only two years more than one out of every ten individuals with a predicted probability of over 0.5 became diagnosed diabetics. The number of individuals on each predicted probability decile provides a benchmark for the impact of a policy proposal that includes our risk-profiling process as a key tool.

Two promising implications can be pinpointed from these results: first, since DM-2 is a silent disease which usually goes un-diagnosed, the analysis here proposed could be leveraged as a stratification tool for prioritizing the application of confirmatory tests on those individual who the model identifies as risk-prone. Second, un-diagnosed individuals subjected to a capillary test can be directed to a priority program that directs them immediately to the laboratory test to be diagnosed.

Here we focus on the first idea since we have the means to estimate its impact. To see this we must consider who is being tested according to present procedures. In table 10 we show the confirmatory test distributions and confirmation rates for each year among the predicted probability deciles.

Table 10
Confirmatory tests by deciles*

Classification	2012		2013		2014	
	Confirmatory tests performed (percent of total tests)	**Positive confirmatory tests (percent of confirmation)**	**Confirmatory tests performed (percent of total tests)**	**Positive confirmatory tests (percent of confirmation)**	**Confirmatory tests performed (percent of total tests)**	**Positive confirmatory tests (percent of confirmation)**
Pr<0.1	28,628 (44%)	4,580 (16%)	31,651 (45%)	4,983 (16%)	33,454 (46%)	4,957 (15%)
0.1≤Pr<0.2	11,538 (18%)	2,737 (24%)	12,729 (18%)	3,015 (24%)	12,774 (18%)	2,875 (23%)
0.2≤Pr<0.3	5,529 (9%)	1,688 (31%)	6,114 (9%)	1,864 (31%)	6,315 (9%)	1,829 (29%)
0.3≤Pr<0.4	5,606 (9%)	1,977 (35%)	6,406 (9%)	2,264 (35%)	6,576 (9%)	2,168 (33%)
0.4≤Pr<0.5	5,705 (9%)	2,117 (37%)	6,071 (9%)	2,356 (39%)	6,047 (9%)	2,217 (37%)
0.5≤Pr<0.6	3,604 (6%)	1,580 (44%)	3,747 (5%)	1,629 (43%)	3,683 (5%)	1,538 (42%)
0.6≤Pr<0.7	2,042 (3%)	1,092 (54%)	2,009 (3%)	1,047 (52%)	1,964 (3%)	1,027 (52%)
0.7≤Pr<0.8	1,276 (2%)	800 (63%)	1,367 (2%)	895 (47%)	1,163 (2%)	745 (64%)
0.8≤Pr	564 (1%)	405 (72%)	578 (1%)	402 (70%)	506 (1%)	352 (70%)
Total tests	64,492 (100%)	19,676 (31%)	70,672 (100%)	18,455 (26%)	72,482 (100%)	17,708 (24%)

* On this table we report the number of confirmatory tests reported per estimated probability decile for each year, percentage in parenthesis, as well as the number of this tests that resulted in diagnosis, percentage in parenthesis.

From the results displayed on the table we observe that the higher the estimated probability the bigger the chance a rights-holder has of confirming diagnosis whenever a confirmatory test is performed. It is important to recall that the capillary glucose test is used as a filter to be assigned to a confirmatory laboratory test, and hence each individual reported on the table allegedly had an elevated capillary glucose measurement. Therefore, the difference in probability of confirmation can be interpreted as the contribution that the estimation has on the Institute's capacity to provide confirmatory diagnosis, when the number of confirmatory tests is fixed or predetermined.

For example, if the 33,454 confirmatory tests performed during 2014 on individuals with a predicted probability smaller than 0.1 had been applied on individuals with a predicted risk of 0.6 or more, the mean confirmation rate would have been 58%, instead of 15%. In other words, the alternative procedure would have rendered 14,446 more confirmed diagnosis from applying the same number of laboratory tests. Moreover, if Institute's procedures led to apply all 2014 confirmatory tests on a population predefined by an estimation of at least 0.5 risk, 18,537 more diabetics would have been diagnosed performing the same 65,166 lab tests that were applied on a population for which our model indicates a risk smaller than 0.5 of being positively diagnosed. Had the tests been applied on individuals with risk of over 0.5 for the whole period 2012-2014, our calculations indicate that the Institute would have detected *over 50 thousand additional positive cases for diabetes, a 90% increase when compared to the 55 thousand reported diagnosis between 2012 and 2014*.

Also, we can see that the number of confirmatory tests is smaller than the new diagnosed population, this, we believe, is conditional on current procedures, as doctors tend to under-register since they need to type the confirmatory test as an extra category on their diagnosis report which does not necessarily have to exist in order for the test to be performed. In other words, assuming that omitted confirmatory tests are distributed uniformly across the population, the impact of leveraging the risk-profiling methodology hereby introduced to detect and direct to lab confirmatory tests would be considerably larger, around 6 times more (300,000 new diagnosis).

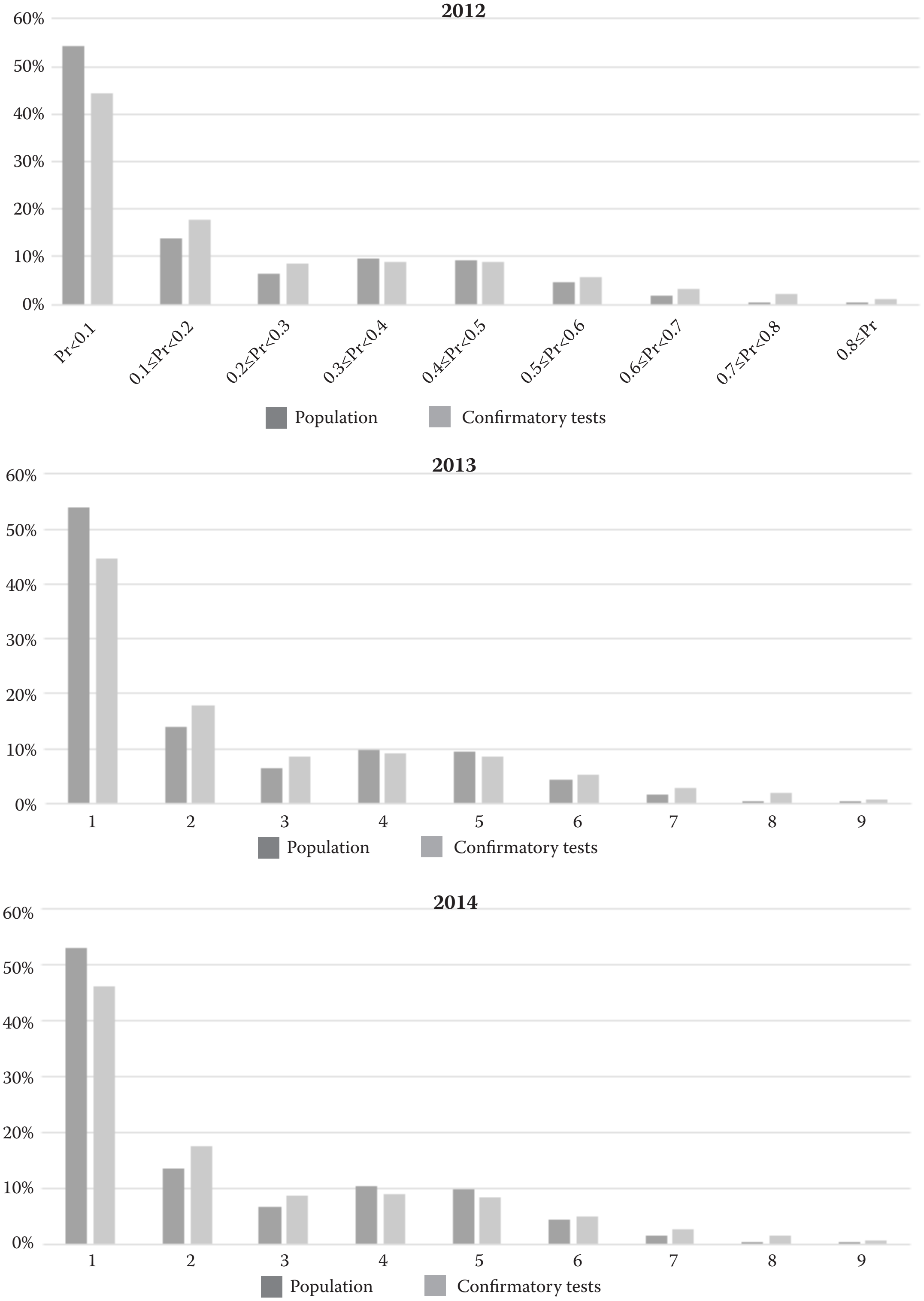
This point is further supported by the similarity between the population distribution among deciles and the distribution of confirmatory tests. To illustrate this we show in figure 12 a bar graph per year with the percent of population that each decile has along with the percentage of confirmatory tests it received.

The figure clearly illustrates the similitude between the percentage of population and tests by decile. Even though there is a smaller concentration of tests relative to the population in the first decile, there are a higher percentage of tests in the second decile. Overall, between 93-94% of each year's population has an estimated probability of under 0.5 and between 88-90% of the tests being performed on them.

Finally, since IMSS has updated information on each worker, we can estimate the risk of being diagnosed with DM-2 for every beneficiary of the system by solely asking them to supply us with the answers to 4 specific questions:

a) social security number and medical aggregate,
b) weight,
c) height and
d) blood pressure.

Figure 12
Population vs. confirmatory test distribution



With a hold on this information we would be able to obtain the rest of the characteristics that lead to a complete profile from leveraging IMSS databases and thus predict individual risk accurately enough.

On the next section we draw some concluding remarks and discuss how the previous results can be utilized in order to significantly improve the capacity of IMSS preventive health apparatus so that it can detect pathological cases more efficiently.

Concluding remarks

This study introduces a methodology that can be applied to estimate the probability of being diagnosed with Type 2 Diabetes, DM-2. The econometric model here presented can be used to calculate the risk profile of beneficiaries of Mexico's largest public health care provider, IMSS. A logit model calibrated for three subsequent years lends itself to not only evaluate the statistical stability of estimated coefficients across periods, but also to evaluate its overall predictive power in a dynamic context.

As the methodology hereby introduced lends itself to extrapolation to out of sample individuals, it enhances its applicability for risk detection for the population at large. More specifically, the latter feature of the model permits us to propose a scheme based on a simple four-point questionnaire that can potentially be implemented at preventive care entry-points in order to effectively detect, channelize and follow-up risk-prone subjects.

As described from the outset, chronic-degenerative diseases like DM-2 can impact several aspects of a person's life; from low productivity due to poor health to living with disabilities and even premature death. Recognizing that early prevention capabilities are key as they can accelerate diagnoses so that disease-prone individuals can act upon pathological conditions in order to prevent costly hospitalizations, the econometric methodology introduced here can be a powerful device in the hands of a public healthcare provider like IMSS.

To date, Mexico is by all accounts suffering from the effects that a lack of suitable detection strategies can ensue, as the country presents a picture of obesity and diabetes prevalence that places it at levels close to an epidemiologic crisis. Therefore, the provision of effective policy measures to detect, channelize, diagnose and follow-up with preventive health information and advice is today a priority for the nation's health authorities. We thus hope that the results hereby presented contribute in some way or another to moving the ball forward in order to open the path for initiatives that attack this problem more aggressively.

Our expectation is that the conclusions reached in this study constitute valuable information and offer insights that can aid in improving detection policy for chronic diseases in Mexico.

References

Akter, S., Rahman, M. M., Abe, S. K., & Sultana, P. (2014). Prevalence of diabetes and prediabetes and their risk factors among Bangladeshi adults: a nationwide survey. *Bulletin of the World Health Organization*, *92*(3), 204-213A.

Barraza-Lloréns M., Guajardo-Barrón V., Picó J., García R., Hernández C., Mora F., Athié J., Crable E., Urtiz A. (2015). *Carga económica de la diabetes mellitus en México*, 2013. México: Funsalud.

Bergtold, J., Yeager, E., & Featherstone, A. (2011). *Sample size and robustness of Inferences from Logistic Regression in the presence of Nonlinearity and Multicollinearity*. Agricultural and Applied Economics Association's.

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology,158*(3), 280-287.

Dietrich, J. K., & Sorensen, E. (1984). An application of logit analysis to prediction of merger targets. *Journal of Business Research*, *12*(3), 393-402.

Feigin, V. L., Forouzanfar, M. H., Krishnamurthi, R., Mensah, G. A., Connor, M., Bennett, D. A., ... & Murray, C. (2014). Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010) and the GBD Stroke Experts Group. Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. *Lancet*, *383*(9913), 245-54.

Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., & Shaw, J. E. (2014). Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, *103*(2), 137-149.

Hosmer Jr., D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

IMSS (2014). *Informe al Ejecutivo Federal y al Congreso de la Unión sobre la situación financiera y los riesgos del Instituto Mexicano del Seguro Social 2013-2014.* México: Instituto Mexicano del Seguro Social. http://www.imss.gob.mx/sites/all/statics/pdf/informes/20132014/21_InformeCompleto. pdf

INSP y SS (2012). *Encuesta Nacional de Salud y Nutrición 2012. Evidencia para la política pública en salud. Diabetes Mellitus: la urgencia de reforzar la respuesta en políticas públicas para su prevención y control.* México: Instituto Nacional de Salud Pública/Secretaría de Salud. http://ensanut.insp.mx/ doctos/analiticos/DiabetesMellitus.pdf

International Diabetes Federation (2014). *International Diabetes Foundation Atlas 2014.* http://www.idf.org/diabetesatlasLee, S. (2005). Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data.*International Journal of Remote Sensing*, *26*(7), 1477-1491.

Lee, S. (2005). Application of Logistic Regression Model and Its Validation for Landslide Susceptibility Mapping Using GIS and Remote Sensing Data. *International Journal of Remote Sensing, 26*(7), 1477-1491.

Martin, D. (1977). Early warning of bank failure: A logit regression approach.*Journal of banking & finance*, *1*(3), 249-276.

Narayan, K. V., Boyle, J. P., Thompson, T. J., Sorensen, S. W., & Williamson, D. F. (2003). Lifetime risk for diabetes mellitus in the United States. *Jama*, *290*(14), 1884-1890.

OECD (2015). *OECD Reviews of Health Systems: Mexico 2015.* Organization for Economic Cooperation and Development.

Secretaría de Salud (marzo, 2011). *Programa Emergente 2011-2012. Prevención y Control del Sobrepeso y Obesidad* (Documento de trabajo). México: Secretaría de Salud. http://www.salud.gob.mx/unidades/cdi/pot/fxi/CENAPRECE/PROG2011_2012.pdf

Stevens, R. J., Kothari, V., Adler, A. I., Stratton, I. M., Holman, R. R., & United Kingdom Prospective Diabetes Study (UKPDS) Group. (2001). The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clinical Science*, *101*(6), 671-679.

Tapia-Conyer, R., Gallardo-Rincón, H., & Saucedo-Martinez, R. (2013). CASALUD: an innovative health-care system to control and prevent non-communicable diseases in Mexico. *Perspectives in public health*, 1757913913511423.

Valenzuela, T. D., Roe, D. J., Cretin, S., Spaite, D. W., & Larsen, M. P. (1997). Estimating effectiveness of cardiac arrest interventions a logistic regression survival model. *Circulation*, *96*(10), 3308-3313.

Villalpando, S., De la Cruz, V., Rojas, R., Shamah-Levy, T., Ávila, M. A., Gaona, B., ... & Hernández, L. (2010). Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population: a probabilistic survey. *Salud pública de México*, *52*, S19-S26.

Whiting, D. R., Guariguata, L., Weil, C., & Shaw, J. (2011). IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes research and clinical practice*, *94*(3), 311-321.